РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

СООБЩЕНИЯ ПО ПРИКЛАДНОЙ МАТЕМАТИКЕ

ВОРОНЦОВ К. В., ФРЕЙ А. И., ТОЛСТИХИН И. О.

КОМБИНАТОРНЫЕ ОЦЕНКИ ВЕРОЯТНОСТИ ПЕРЕОБУЧЕНИЯ: ТЕОРЕТИКО-ГРУППОВОЙ ПОДХОД

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН
МОСКВА, 2011

УДК 004.852

Ответственный редактор
чл.-корр. РАН К. В. Рудаков

В рамках комбинаторной теории переобучения развивается теоретико-групповой подход, позволяющий получать точные оценки вероятности переобучения для симметричных и рандомизированных семейств алгоритмов.

Ключевые слова: *теория статистического обучения, обобщающая способность, вероятность переобучения, теория групп*
.

Рецензенты: В. В. Стрижов,
А. Г. Дьяконов

Научное издание

Содержание

1. Проблема переобучения при восстановлении зависимостей по эмпирическим данным

При решении задач распознавания образов, восстановления регрессии, прогнозирования всегда возникает проблема выбора по неполной информации. Имея лишь конечную обучающую выборку объектов, требуется из заданного множества алгоритмов выбрать алгоритм, который ошибался бы как можно реже не только на объектах наблюдаемой обучающей выборки, но и на объектах скрытой контрольной выборки, которая в момент выбора алгоритма ещё неизвестна. Если частота ошибок на контрольной выборке оказывается значительно выше, чем на обучающей, то говорят, что произошло «переобучение» (overtraining) или «переподгонка» (overfitting) алгоритма — он слишком хорошо описывает конкретные данные, но не обладает способностью к обобщению этих данных, не восстанавливает порождающую их зависимость и не пригоден для построения прогнозов.

На практике обучающая выборка формируется раньше, чем контрольная. Таким образом, обучающая и контрольная выборки могут иметь различные статистические свойства. Низкое качество на контроле может быть обусловлено не только переобучением алгоритма, но и нестационарностью данных во времени.

Процесс формирования обучающей выборки также является важным фактором, влияющим на переобучение. В частности, большое значение имеет представительность (репрезентативность) обучающей выборки. Данная проблема хорошо известна при проведении социалогических опросов. Как ограничить круг респондетнов, чтобы тем не менее представить весь спектр общественного мнения? Аналогичный вопрос возникает и при форми-

ровании обучающей выборки.

В дальнейшем выборка данных будет предполагаться репрезентативной и стационарной. Главной целью ставится исследование свойств метода обучения как такового.

На практике переобучение оценивается количественно с помощью процедуры скользящего контроля (кросс-валидации). Фиксируется некоторое множество разбиений исходной выборки на две подвыборки — обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. *Оценкой скользящего контроля* называется средняя по всем разбиениям величина ошибки на контрольных подвыборках. Аналогичным образом определяется и *оценка вероятности переобучения* — это доля разбиений, при которых средняя ошибка на контрольной подвыборке превышает среднюю ошибку на обучающей подвыборке более чем на заданную величину ε. Главный недостаток данного подхода — большая вычислительная сложность, связанная с многократной настройкой алгоритма классификации. Точность оценки скользящего контроля зависит от стабильности метода обучения и от числа разбиений, по которым производилось усреднение. В комбинаторной теории переобучения рассматривается множество всех возможных разбиений и ставится задача получения вычислительно эффективных формул для оценок скользящего контроля и вероятности переобучения [**?**, **?**].

Теоретико-групповой подход [**?**, **?**, **?**] позволяет получать такие формулы для случаев, когда семейство алгоритмов обладает некоторой симметрией, а методом обучения является *рандомизированная минимизация эмпирического риска.* Рандомизация означает, что если в семействе существует несколько алгоритмов, допускающих одинаковое минимальное число ошибок на обучающей выборке, то из них равновероятно выбирается любой.

1.1. Задача обучения по прецедентам

Пусть задана генеральная выборка $\mathbb{X} = (x_1, \ldots, x_L)$, состоящая из L объектов. Произвольный алгоритм классификации, примененный к данной выборке, порождает бинарный вектор ошибок $a \equiv \big(I(a, x_i)\big)_{i=1}^{L}$, где $I(a, x_i) \in \{0, 1\}$ — индикатор ошибки алгоритма a на объекте x_i. В дальнейшем алгоритмы будут отождествляться с векторами их ошибок на выборке $\mathbb{X}$.

Обозначим через $\mathbb{A} = \{0, 1\}^L$ множество всех возможных векторов ошибок длины L. Через $[\mathbb{X}]^\ell$ обозначим множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$. *Число ошибок* алгоритма a на выборке $U \subseteq \mathbb{X}$ обозначим через $n(a, U) = \sum\limits_{x \in U} I(a, x)$. Величину $\nu(a, U) = n(a, U)/|U|$ будем называть *частотой ошибок* алгоритма a на выборке U. *Уклонение частот* на разбиении $\mathbb{X} = X \sqcup \bar{X}$ определим как разность частот ошибок на контроле и на обучении: $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

Пусть $A \subset \mathbb{A}$ — множество алгоритмов с попарно различными векторами ошибок. Обозначим через $A(X)$ множество алгоритмов с минимальным числом ошибок на обучающей выборке X:

$$A(X) = \arg\min_{a \in A} n(a, X). \tag{1}$$

Частоту ошибок на обучающей выборке называют *эмпирическим риском. Минимизация эмпирического риска* μ — это метод обучения, который из заданного множества $A \subset \mathbb{A}$ выбирает алгоритм $a \in A$, допускающий наименьшее число ошибок на обучающей выборке X. Таким образом, для всех $X \in [\mathbb{X}]^\ell$ выполнено $\mu X \in A(X)$.

В следующей таблице показан пример, когда минимизация эмпирического риска приводит к переобучению. Столбцы таблицы соответствуют алгоритмам, строки — объектам обучающей выбор-

ки $\{x_1, \ldots, x_\ell\}$ и контрольной выборки $\{x_{\ell+1}, \ldots, x_L\}$.

$$\begin{array}{c} \\ x_1 \\ \cdots \\ x_\ell \\ \\ x_{\ell+1} \\ \cdots \\ x_L \end{array} \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & \cdots & a_d & \cdots & a_D \end{array} \\ \left(\begin{array}{cccccc} 0 & 1 & \cdots & 0 & \cdots & 1 \\ 1 & 1 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \hline 1 & 1 & \cdots & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & \cdots & 0 \end{array} \right) \end{array}$$

В данном примере переобучение могло быть следствием «неудачного» разбиения генеральной выборки на обучение и контроль. Поэтому вводится функционал *вероятности переобучения*, равный доле разбиений выборки, при которых возникает переобучение [**?**, **?**]:

$$Q_{\mu,\varepsilon}(A) = \mathbb{E}[\delta(\mu X, X) \geq \varepsilon], \text{ где } \mathbb{E} = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell}.$$

Тут и далее квадратные скобки — нотация Айверсона [**?**], переводящая логическое выражение в число 0 или 1 по правилам [истина] = 1, [ложь] = 1.

Для краткости в тех случаях, когда из контекста ясно, о каком методе обучения идет речь, мы будем опускать индекс μ и писать просто $Q_\varepsilon(A)$.

Функционал $Q_\varepsilon(A)$ уже не зависит от выбора разбиения и характеризует качество данного метода обучения на данной генеральной выборке.

При минимизации эмпирического риска может возникать неоднозначность — несколько алгоритмов из $A(X)$ могут иметь одинаковое число ошибок на обучающей выборке. В [**?**] для устранения неоднозначности и получения точных верхних оценок вероятности переобучения использовалась *пессимистичная* минимизация эмпирического риска — предполагалось, что в случае неоднозначности выбирается алгоритм с наибольшим числом ошибок

на генеральной выборке $\mathbb{X}$. Это не устраняет неоднозначность окончательно. Возможны ситуации, когда несколько алгоритмов имеют наименьшее число ошибок на обучающей выборке X и одинаковое число ошибок на генеральной выборке $\mathbb{X}$. В таких случаях на множестве алгоритмов вводился линейный порядок, и среди неразличимых алгоритмов выбирался алгоритм с бóльшим порядковым номером. Введение приоритетности алгоритмов является искусственным приемом, не имеющим адекватных аналогов среди известных методов обучения.

1.2. Рандомизированная минимизация эмпирического риска

Рандомизированный метод минимизации эмпирического риска выбирает произвольный алгоритм из множества $A(X)$ случайно и равновероятно [**?**]. Поскольку в задаче статистического обучения появляется второй независимый источник случайности, определение вероятности переобучения $Q_\varepsilon(A)$ приходится модифицировать. Наиболее естественный вариант модификации — усреднение по множеству $A(X)$:

$$Q_\varepsilon(A) = \mathbb{E}\frac{1}{|A(X)|}\sum_{a\in A(X)}\big[\delta(a,X)\geq\varepsilon\big]. \tag{2}$$

Переставим местами знаки суммирования, $\mathbb{E}\sum = \sum\mathbb{E}$. Получим сумму по всем алгоритмам $a \in A$, каждое слагаемое которой обозначим через $Q_\varepsilon(a,A)$ и назовём *вкладом алгоритма* a в вероятность переобучения:

$$Q_\varepsilon(A) = \sum_{a\in A} Q_\varepsilon(a,A), \quad Q_\varepsilon(a,A) = \mathbb{E}\frac{\big[a\in A(X)\big]}{|A(X)|}\big[\delta(a,X)\geq\varepsilon\big].$$

Аналогичным образом введём *вероятность реализации алгоритма* a:

$$P(a,A) = \mathbb{E}\frac{\big[a\in A(X)\big]}{|A(X)|}. \tag{3}$$

В частном случае, когда множество A состоит из единственного алгоритма, вероятность переобучения (2) и вероятность реализации (3) принимают привычный вид:

$$Q_\varepsilon(A) = \mathbb{E}\big[\delta(\mu X, X) \geq \varepsilon\big], \quad P(a, A) = \mathbb{E}\big[a = \mu X\big].$$

Рандомизированный метод минимизации эмпирического риска μ_r занимает промежуточное положение между *оптимистичным* и *пессимистичным* методами:

$$\mu_o X = \arg\min_{a \in A(X)} n(a, \bar{X}) \text{ — оптимистичный ММЭР;}$$
$$\mu_p X = \arg\max_{a \in A(X)} n(a, \bar{X}) \text{ — пессимистичный ММЭР.}$$

Приводим без доказательства следующее утверждение: для произвольного множества алгоритмов $A \subseteq \mathbb{A}$ и каждого $\varepsilon \in (0, 1]$ справедлива цепочка неравенств:

$$Q_{\mu_o, \varepsilon}(A) \leq Q_{\mu_r, \varepsilon}(A) \leq Q_{\mu_p, \varepsilon}(A). \tag{4}$$

1.3. Перестановки объектов

Введём симметрическую группу S_L всех $L!$ перестановок, действующую на выборке $\mathbb{X} = (x_1, \ldots, x_L)$. Возьмём произвольную перестановку $\pi \in S_L$. Обозначим через πx тот объект, в который объект $x \in \mathbb{X}$ переходит под действием перестановки π. Действие перестановок на объектах естественным образом переносится на подмножества объектов, на алгоритмы как бинарные векторы ошибок длины L и на множества алгоритмов:

действие перестановки π на подмножество объектов:

$$\pi X = \big\{\pi x \colon x \in X\big\};$$

действие перестановки π на алгоритм:

$$\pi \boldsymbol{a} = \big(I(\pi a, x_i)\big)_{i=1}^L = \big(I(a, \pi^{-1} x_i)\big)_{i=1}^L;$$

действие перестановки π на множество алгоритмов:

$$\pi A = \{\pi a \colon a \in A\}.$$

Заметим, что действие одной и той же перестановки π сначала на выборку $\mathbb{X}$, затем на алгоритм a, восстанавливает исходный вектор ошибок алгоритма a. Благодаря такому определению действие на алгоритм обладает рядом полезных свойств.

Лемма 1. *Свойства действия произвольной перестановки* $\pi \in S_L$*:*

1) $I(\pi a, \pi x) = I(a, x)$ *для любых* $a \in A$ *и* $x \in \mathbb{X}$*;*

2) $n(\pi a, \mathbb{X}) = n(a, \mathbb{X})$ *для любого* $a \in A$*;*

3) $n(\pi a, \pi X) = n(a, X)$ *для любых* $a \in A$ *и* $X \subseteq \mathbb{X}$*;*

4) $\delta(\pi a, \pi X) = \delta(a, X)$ *для любых* $a \in A$ *и* $X \subseteq \mathbb{X}$*;*

5) $[a \in A(X)] = [\pi a \in (\pi A)(\pi X)]$ *для любых* $a \in A$ *и* $X \subseteq \mathbb{X}$*;*

6) $|A(X)| = |(\pi A)(\pi X)|$ *для любых* A *и* $X \subseteq \mathbb{X}$*;*

7) $\rho(a, a') = \rho(\pi a, \pi a')$ *для любых* $a, a' \in A$*, где* $\rho(a, a')$ *— расстояние Хэмминга векторами ошибок алгоритмов* a *и* a'*:*

$$\rho(a, a') = \sum_{x \in \mathbb{X}} |I(a, x) - I(a', x)|.$$

Доказательство. Свойство 1) следует из определения:

$$I(\pi a, \pi x) = I(a, \pi^{-1}\pi x) = I(a, x).$$

Свойство 2) следует из свойства 1):

$$n(\pi a, \mathbb{X}) = \sum_{i=1}^{L} I(\pi a, x_i) = \sum_{i=1}^{L} I(\pi a, \pi x_i) = \sum_{i=1}^{L} I(a, x_i) = n(a, \mathbb{X}).$$

Свойство 3) также следует из свойства 1):

$$n(\pi a, \pi X) = \sum_{x \in \pi X} I(\pi a, x) = \sum_{x \in X} I(\pi a, \pi x) = \sum_{x \in X} I(a, x) = n(a, X).$$

Свойство 4) следует из свойства 3):

$$\delta(a,X)=\frac{L-n(a,X)}{k}-\frac{n(a,X)}{\ell}=$$
$$=\frac{L-n(\pi a,\pi X)}{k}-\frac{n(\pi a,\pi X)}{\ell}=\delta(\pi a,\pi X).$$

Свойство 5) следует из определения 1 и свойства 1):

$$a_0\in A(X)\ \Leftrightarrow\ a_0\in\arg\min_{a\in A}n(a,X)\ \Leftrightarrow$$
$$\forall a\in A\to n(a_0,X)\le n(a,X)\ \Leftrightarrow$$
$$\forall a\in A\to n\big(\pi a_0,\pi X\big)\le n\big(\pi a,\pi X\big)\ \Leftrightarrow$$
$$\forall a'\in \pi A\to n\big(\pi a_0,\pi X\big)\le n\big(a',\pi X\big)\ \Leftrightarrow$$
$$\pi a_0\in\arg\min_{a'\in\pi A}n(a',\pi X)\ \Leftrightarrow\ \pi a_0\in(\pi A)(\pi X).$$

Свойство 6) следует из свойства 5):

$$|A(X)|=\sum_{a\in A}[a\in A(X)]=\sum_{a\in A}[\pi a\in(\pi A)(\pi X)]=$$
$$=\sum_{a'\in\pi A}[a'\in(\pi A)(\pi X)]=|(\pi A)(\pi X)|.$$

Свойство 7) следует из свойства 1):

$$\rho(\pi a,\pi a')=\sum_{x\in\mathbb{X}}|I(\pi a,x)-I(\pi a',x)|=$$
$$=\sum_{x'\in\mathbb{X}}|I(\pi a,\pi x')-I(\pi a',\pi x')|=$$
$$=\sum_{x'\in\mathbb{X}}|I(a,x')-I(a',x')|=\rho(a,a').$$

■

1.4. Группа симметрий множества алгоритмов.

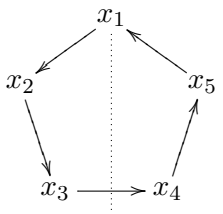
Рассмотрим множество всех перестановок, действие которых на множество A не меняет его:

$$\operatorname{Sym} A = \{\pi \in S_L \colon \pi A = A\}.$$

Если подействовать любой из перестановок $\pi \in \operatorname{Sym} A$ на строки матрицы ошибок множества A, то получится ровно то же самое множество столбцов; переставив столбцы, можно получить исходную матрицу ошибок. Очевидно, множество $\operatorname{Sym} A$ является группой. Будем называть её *группой симметрий* множества алгоритмов A.

Пример 1. *Рассмотрим множество алгоритмов, заданное матрицей ошибок*

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} a_1 & a_2 & a_3 & a_4 & a_5 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$



Группа симметрий данного множества алгоритмов совпадает с группой симметрий правильного пятиугольника и называется диэдральной группой. *Образующими элементами группы являются циклическая перестановка* $\pi_1 = (x_1, x_2, x_3, x_4, x_5)$ *и осевая симметрия* $\pi_2 = (x_2, x_5)(x_3, x_4)$.

Пусть далее $G \subseteq \operatorname{Sym} A$ — произвольная подгруппа группы $\operatorname{Sym} A$.

Для любой перестановки $\pi \in G$ и любого алгоритма $a \in A$ алгоритм πa снова лежит в A. В таких случаях говорят, что группа G *действует* на множестве A.

Орбитой алгоритма $a \in A$ называется множество алгоритмов $Ga = \{\pi a \colon \pi \in G\}$. Орбита также целиком лежит в A. Орбиты

двух различных алгоритмов Ga и Ga' либо совпадают, либо не пересекаются. Следовательно, множество A разбивается на непересекающиеся подмножества — орбиты:

$$A = \bigsqcup_{\omega \in \Omega(A)} \omega = \bigsqcup_{\omega \in \Omega(A)} Ga_\omega,$$

где $\Omega(A)$ — множество всех орбит в A, a_ω — произвольный представитель орбиты ω.

Из свойства 2) леммы 1 следует, что алгоритмы одной орбиты обязательно лежат в одном слое. Обратное, вообще говоря, неверно.

Лемма 2. *Алгоритмы из одной орбиты имеют равные вероятности реализации и равные вклады в вероятность переобучения: для любой перестановки π из G*

$$P(a, A) = P(\pi a, A), \quad Q_\varepsilon(a, A) = Q_\varepsilon(\pi a, A).$$

Доказательство. Воспользуемся определением вероятности реализации (3), свойствами 5), 6) из леммы 1, и свойством $A = \pi A$:

$$P(a, A) = \mathbb{E}\frac{\big[a \in A(X)\big]}{|A(X)|} = \mathbb{E}\frac{\big[\pi a \in (\pi A)(\pi X)\big]}{|(\pi A)(\pi X)|} = \mathbb{E}\frac{\big[\pi a \in A(\pi X)\big]}{|A(\pi X)|}.$$

Под знаком $\mathbb{E}$ можно всюду заменить πX на X, так как результат не зависит от порядка суммирования разбиений:

$$P(a, A) = \mathbb{E}\frac{\big[\pi a \in A(X)\big]}{|A(X)|} = P(\pi a, A).$$

Воспользуемся определением вероятности реализации (3),

свойствами 4), 5), 6) из леммы 1, и свойством $A = \pi A$:

$$\begin{aligned} Q_\varepsilon(a, A) &= \mathbb{E}\frac{\big[a \in A(X)\big]}{|A(X)|}\big[\delta(a, X) \geq \varepsilon\big] = \\ &= \mathbb{E}\frac{\big[\pi a \in (\pi A)(\pi X)\big]}{|(\pi A)(\pi X)|}\big[\delta(\pi a, \pi X) \geq \varepsilon\big] = \\ &= \mathbb{E}\frac{\big[\pi a \in A(\pi X)\big]}{|A(\pi X)|}\big[\delta(\pi a, \pi X) \geq \varepsilon\big]. \end{aligned}$$

Вновь заменяя πX на X под знаком $\mathbb{E}$, получим:

$$Q_\varepsilon(a, A) = \mathbb{E}\frac{\big[\pi a \in A(X)\big]}{|A(X)|}\big[\delta(\pi a, X) \geq \varepsilon\big] = Q_\varepsilon(\pi a, A).$$

■

Разложение вероятности переобучения по орбитам множества алгоритмов. Из теоремы о равном вкладе алгоритмов одной орбиты немедленно следует формула разложения вероятности переобучения по орбитам. Она является основным инструментом получения точных оценок для рандомизированного метода минимизации эмпирического риска.

Теорема 1. *Для любой генеральной выборки $\mathbb{X}$, любого множества алгоритмов A с попарно различными векторами ошибок и любого $\varepsilon \in [0, 1]$ справедлива формула разложения вероятности переобучения по орбитам множества A:*

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} |\omega|\, \mathbb{E}\frac{\big[a_\omega \in A(X)\big]}{|A(X)|}\big[\delta(a_\omega, X) \geq \varepsilon\big], \tag{5}$$

где $\Omega(A)$ — множество всех орбит в A, a_ω — произвольный представитель орбиты ω.

Доказательство. Перегруппируем слагаемые в (2) по орбитам множества A, затем применим лемму 2 о равном вкладе алгоритмов одной орбиты:

$$Q_\varepsilon(A) = \sum_{\omega\in\Omega(A)} \sum_{a\in\omega} \mathbb{E} \frac{\big[a \in A(X)\big]}{|A(X)|} \big[\delta(a, X) \geq \varepsilon\big] =$$

$$= \sum_{\omega\in\Omega(A)} |\omega|\, \mathbb{E} \frac{\big[a_\omega \in A(X)\big]}{|A(X)|} \big[\delta(a_\omega, X) \geq \varepsilon\big]. \qquad \blacksquare$$

Разложение вероятности переобучения по орбитам множества разбиений. В некоторых случаях удобнее делать группировку слагаемых не по орбитам множества алгоритмов, а по орбитам множества разбиений. Напомним, что через $[\mathbb{X}]^\ell$ мы обозначаем множество всех ℓ-элементных подмножеств генеральной выборки $\mathbb{X}$.

Представим вероятность переобучения в виде суммы вкладов разбиений:

$$Q_\varepsilon(A) = \mathbb{E}\, Q_\varepsilon(X, A), \quad Q_\varepsilon(X, A) = \frac{1}{|A(X)|} \sum_{a\in A(X)} \big[\delta(a, X) \geq \varepsilon\big],$$

где $Q_\varepsilon(X, A)$ — вклад разбиения $X \sqcup \bar{X}$ в вероятность переобучения. Поскольку разбиениям $X \sqcup \bar{X}$ взаимно однозначно соответствуют выборки $X \in [\mathbb{X}]^\ell$, далее будем говорить также о *вкладе выборки* X в вероятность переобучения.

Орбитой выборки $X \in [\mathbb{X}]^\ell$ называется множество выборок $GX = \big\{\pi X \colon \pi \in G\big\}$. Множество всех выборок длины ℓ разбивается на непересекающиеся орбиты:

$$[\mathbb{X}]^\ell = \bigsqcup_{\tau\in\Omega[\mathbb{X}]^\ell} \tau = \bigsqcup_{\tau\in\Omega[\mathbb{X}]^\ell} GX_\tau,$$

где $\Omega[\mathbb{X}]^\ell$ — множество всех орбит в $[\mathbb{X}]^\ell$, X_τ — произвольный представитель орбиты τ.

Лемма 3. *Выборки из одной орбиты имеют равные вклады в вероятность переобучения:* $Q_\varepsilon(X, A) = Q_\varepsilon(\pi X, A)$ *для любой перестановки* π *из* G.

Доказательство. Воспользуемся сначала определением вклада выборки, свойствами 4), 5), 6) из леммы 1, и затем свойством $A = \pi A$:

$$\begin{aligned} Q_\varepsilon(X, A) &= \sum_{a \in A(X)} \frac{\big[\delta(a, X) \geq \varepsilon\big]}{|A(X)|} = \\ &= \sum_{a' \in (\pi A)(\pi X)} \frac{\big[\delta(a', \pi X) \geq \varepsilon\big]}{|(\pi A)(\pi X)|} = \\ &= \sum_{a' \in A(\pi X)} \frac{\big[\delta(a', \pi X) \geq \varepsilon\big]}{|A(\pi X)|} = Q_\varepsilon(\pi X, A). \end{aligned}$$

■

Теорема 2. *Для любой генеральной выборки* $\mathbb{X}$*, любого множества алгоритмов* A *с попарно различными векторами ошибок и любого* $\varepsilon \in [0, 1]$ *справедлива формула разложения вероятности переобучения по орбитам множества* $[\mathbb{X}]^\ell$*:*

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega[\mathbb{X}]^\ell} \frac{|\tau|}{|A(X_\tau)|} \sum_{a \in A(X_\tau)} \big[\delta(a, X_\tau) \geq \varepsilon\big], \tag{6}$$

где $\Omega[\mathbb{X}]^\ell$ *— множество всех орбит в* $[\mathbb{X}]^\ell$*,* X_τ *— произвольный представитель орбиты* τ.

Доказательство аналогично доказательству теоремы 1.

В качестве примера применения полученных формул рассмотрим множество $\mathbb{A} = \{0, 1\}^L$, состоящее из всех возможных бинарных векторов ошибок.

Теорема 3. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к множе-*

ству всех алгоритмов $\mathbb{A} = \{0,1\}^L$*, дается формулой:*

$$Q_\varepsilon(\mathbb{A}) = \frac{1}{2^k} \sum_{m=\lceil \varepsilon k \rceil}^{k} C_k^m.$$

Доказательство. Для всех перестановок $\pi \in S_L$ выполнено $\pi\mathbb{A} = \mathbb{A}$. Следовательно, $\operatorname{Sym} A = S_L$. Заметим, что для каждой пары обучающих выборок X, X' возможно указать перестановку $\pi \in S_L$, такую что $X' = \pi X$. Такую ситуацию называют «транзитивным действием группы S_L на множестве $[\mathbb{X}]^\ell$». Для нас это означает, что имеется лишь одна орбита $\tau = [\mathbb{X}]^\ell$. Выбрав произвольную выборку X в качестве ее представителя, и воспользовавшись теоремой 2, получим

$$Q_\varepsilon(\mathbb{A}) = \frac{1}{|\mathbb{A}(X)|} \sum_{a \in \mathbb{A}(X)} [\delta(a, X) \geq \varepsilon].$$

Множество $\mathbb{A}(X)$ состоит из всех алгоритмов, не допускающих ошибок на X. Следовательно, $|\mathbb{A}(X)| = 2^k$. Для завершения доказательства осталось заметить, что переобученными в $\mathbb{A}(X)$ будут те и только те алгоритмы, у которых не менее $\lceil \varepsilon k \rceil$ ошибок на контрольной выборке. ■

Из сформулированных выше теорем 1 и 2 о двух видах разложения вероятности переобучения по орбитам действия группы симметрии легко следует следующее разложение, объединяющее оба вида:

Теорема 4. *Для любой генеральной выборки* $\mathbb{X}$*, любого множества алгоритмов* A *с попарно различными векторами ошибок и любого* $\varepsilon \in [0,1]$ *справедлива формула разложения вероятности переобучения по орбитам множества* $[\mathbb{X}]^\ell$*:*

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} \frac{|\omega|}{C_L^\ell} \sum_{\tau \in \Omega[\mathbb{X}]^\ell} |\{X \in \tau \colon a_\omega \in A(X)\}| \frac{\left[\delta(a_\omega, X_\tau) \geq \varepsilon\right]}{|A(X_\tau)|}.$$

где $\Omega[\mathbb{X}]^\ell$ *— множество всех орбит в* $[\mathbb{X}]^\ell$, $\Omega(A)$ *— множество всех орбит в* A, a_ω *— произвольный представитель орбиты* ω, X_τ *— произвольный представитель орбиты* τ.

Этим разложением мы воспользуемся позже при выводе точной оценки вероятности переобучения для хэммингова шара.

Завершая параграф, интересно рассмотреть частный случай, когда все алгоритмы из A имеют равное число ошибок на полной выборке.

Следствие 1. *Пусть все* $a \in A$ *имеют равное число ошибок на полной выборке:* $n(a, \mathbb{X}) = m$. *Тогда вероятность переобучения рандомизированного метода минимизации эмпирического риска записывается в виде*

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \left[\min_{a \in A} n(a, X_\tau) \leq \frac{\ell}{L}(m - \varepsilon k) \right]. \qquad (7)$$

Доказательство. Отметим, что в рассматриваемом случае для любой обучающей выборки X все алгоритмы из множества $A(X)$ либо переобучены, либо нет. Действительно, согласно определению $A(X)$ они имеют равное число ошибок на обучении. Число ошибок на полной выборке одинаково поскольку в силу специфики рассматриваемого случая все алгоритмы из A лежат в одном слое. Следовательно, все алгоритмы из A имеют равное число ошибок на контрольной выборке, и равные уклонения частот. Тогда, применяя формулу (6), получим следующее выражение для вероятности переобучения:

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \left[\delta(a, X_\tau) > \varepsilon \right].$$

Для получения формулы (7) осталось выразить уклонение частот $\delta(a, X_\tau)$ через число ошибок лучшего алгоритма на обучении и количество ошибок на полной выборке m. ■

1.5. Теорема о порождающих и запрещающих объектах

Первый подход, позволивший получать точные оценки вероятности переобучения в рамках слабой вероятностной аксиоматики, основан на выделении порождающих и запрещающих объектов [**?**].

Гипотеза 1. *Пусть множество A, выборка $\mathbb{X}$ и детерминированный метод обучения μ таковы, что для каждого алгоритма $a \in A$ можно указать пару непересекающихся подмножеств $X_a \subset \mathbb{X}$ и $X'_a \subset \mathbb{X}$, удовлетворяющую условию*

$$\big[\mu X{=}a\big] = \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{8}$$

Множество X_a называется *порождающим*, X'_a — *запрещающим* для алгоритма a. Гипотеза 1 означает, что метод μ выбирает алгоритм a тогда и только тогда, когда в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{X}\backslash X_a\backslash X'_a$ называются *нейтральными* для алгоритма a.

Для произвольного алгоритма $a \in A$ введём следующие обозначения:

$L_a = L - |X_a| - |X'_a|$ — число нейтральных объектов в генеральной выборке;

$\ell_a = \ell - |X_a|$ — число нейтральных объектов в обучающей выборке;

$m_a = n(a, \mathbb{X}\backslash X_a\backslash X'_a)$ — число ошибок алгоритма a на нейтральных объектах;

$s_a(\varepsilon) = \frac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_a)$ — наибольшее число ошибок алгоритма a на нейтральных обучающих объектах $X \setminus X_a$, при котором имеет место большое уклонение частот ошибок, $\delta(a, X) \geq \varepsilon$.

Введём функцию гипергеометрического распределения:

$$H_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}.$$

Теорема 5. *Если справедлива гипотеза 1, то вероятность получить в результате обучения алгоритм a равна $P_a(A) = \mathsf{P}[\mu X{=}a] = C_{L_a}^{\ell_a}/C_L^\ell$, вероятность переобучения равна*

$$Q_\varepsilon(A) = \sum_{a\in A} P_a H_{L_a}^{\ell_a,m_a}\big(s_a(\varepsilon)\big).$$

Данный результат позволил получить формулы вероятности переобучения для широкого класса модельных семейств алгоритмов, в частности для монотонных и унимодальных сетей.

Теорема 5 получена для детерминированных методов обучения, для которых результатом обучения является один алгоритм $a \in A$. В случае рандомизированного метода результатом обучения является подмножество $A(X) \subseteq A$. Таким образом, множество алгоритмов A порождает множество подмножеств алгоритмов, получающихся в результате обучения

$$\mathfrak{A} = \big\{A(X)\colon X \in [\mathbb{X}]^\ell\big\}.$$

Гипотеза 2. *Пусть множество A и выборка $\mathbb{X}$ таковы, что для каждого $\alpha \in \mathfrak{A}$ можно указать пару непересекающихся подмножеств $X_\alpha \subset \mathbb{X}$ и $X'_\alpha \subset \mathbb{X}$, удовлетворяющую условию*

$$\big[A(X){=}\alpha\big] = \big[X_\alpha \subseteq X\big]\big[X'_\alpha \subseteq \bar{X}\big],\ \ \forall X \in [\mathbb{X}]^\ell. \tag{9}$$

Следующая теорема является непосредственным обобщением теоремы 5 для рандомизированного метода минимизации эмпирического риска.

Теорема 6. *Если справедлива гипотеза 2, то вероятность переобучения рандомизированного метода минимизации эмпирического риска есть*

$$Q_\varepsilon(A) = \sum_{a\in A}\sum_{\alpha\in\mathfrak{A}} \frac{[a\in\alpha]}{|\alpha|}\frac{C_{L_\alpha}^{\ell_\alpha}}{C_L^\ell} H_{L_\alpha}^{\ell_\alpha,m_\alpha^a}\big(s_\alpha^a(\varepsilon)\big),$$

где введены следующие обозначения:

$$L_\alpha = L - |X_\alpha| - |\bar{X}_\alpha|; \quad \ell_\alpha = \ell - |X_\alpha|;$$
$$m^a_\alpha = n(a, \mathbb{X}\backslash X_\alpha \backslash X'_\alpha);$$
$$s^a_\alpha(\varepsilon) = \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_\alpha).$$

Доказательство. Рассмотрим функционал $Q_\varepsilon(A)$. Введем под знак суммироваия по X два вспомогательных суммирования: первое — по всем $\alpha \in \mathfrak{A}$ при условии $\alpha = A(X)$, второе — по всем значениям s числа ошибок алгоритма a на подвыборке $X\backslash X_\alpha$. Очевидно, значение $Q_\varepsilon(A)$ от этого не измениться:

$$\begin{aligned} Q_\varepsilon(A) &= \mathbb{E} \sum_{a\in A(X)} \frac{1}{|A(X)|}[\delta(a,X) \geq \varepsilon] = \\ &= \mathbb{E} \sum_{\alpha\in\mathfrak{A}} \sum_{a\in\alpha} \frac{[\alpha = A(X)]}{|\alpha|}[\delta(a,X) \geq \varepsilon] = \\ &= \mathbb{E} \sum_{\alpha\in\mathfrak{A}} \sum_{a\in\alpha} \sum_{s=0}^{\ell_\alpha} \frac{[\alpha = A(X)]}{|\alpha|}[n(a, X\backslash X_\alpha) = s][\delta(a,X) \geq \varepsilon]. \end{aligned} \tag{10}$$

Число ошибок алгоритма a на обучающей подвыборке X равно $s + n(a, X_\alpha)$, поэтому отклонение частот выражается в виде

$$\delta(a,X) = \frac{n(a,\mathbb{X}) - s - n(a,X_\alpha)}{k} - \frac{s + n(a,X_\alpha)}{\ell},$$

следовательно

$$[\delta(a,X) \geq \varepsilon] = \big[s \leq \frac{\ell}{L}(n(a,\mathbb{X}) - \varepsilon k) - n(a,X_\alpha)\big] = [s \geq s^a_\alpha(\varepsilon)].$$

Подставим полученное выражение в (10), затем заменим $[\alpha = A(X)]$ правой частью равенства (9) и переставим знаки суммирования (очевидно, $\mathbb{E}$ также можно рассматривать как суммирование):

$$Q_\varepsilon(A) = \tag{11}$$

$$= \sum_{\alpha\in\mathfrak{A}} \sum_{a\in\alpha} \sum_{s=0}^{\ell_a} \frac{1}{|\alpha|} \underbrace{\mathbb{E}\big[X_\alpha \subseteq X\big]\big[X'_\alpha \subseteq \bar{X}\big]\big[n(a, X\backslash X_\alpha) = s\big]}_{N(\alpha,a)} \big[s \le s^a_\alpha(\varepsilon)\big].$$

Выделенное в данной формуле выражение $N(\alpha, a)$ есть доля разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ таких, что множество объектов X_α целиком лежит в X, множество объектов X'_α целиком лежит в $\bar{X}$ и в подвыборку $X\backslash X_\alpha$ длины ℓ_α попадает ровно s объектов, на которых алгоритм a допускает ошибку.

Для наглядности представим вектор ошибок a разбитым на шесть блоков:

$$\boldsymbol{a} = \Big(\underbrace{X_\alpha; \overbrace{1,\ldots,1}^{s}; \underbrace{0,\ldots,0}_{X\backslash X_\alpha}}_{X}; \underbrace{X'_\alpha; \overbrace{1,\ldots,1}^{m^a_\alpha - s}; \underbrace{0,\ldots,0}_{\bar{X}\backslash X'_\alpha}}_{\bar{X}} \Big).$$

Число ошибок алгоритма a на объектах, не попадающих ни в X_α, ни в X'_α, равно m^a_α. Существует $C^s_{m^a_\alpha}$ способов выбрать из них s объектов, которые попадут в $X\backslash X_\alpha$. Для каждого из этих способов имеется ровно $C^{\ell_\alpha - s}_{L_a - m^a_\alpha}$ способов выбрать $\ell_\alpha - s$ объектов, на которых алгоритм a не допускает ошибку, и которые также попадут в $X\backslash X_\alpha$. Тем самым однозначно определяется состав выборки $X\backslash X_\alpha$, а, значит, и состав выборки $\bar{X}\backslash X'_\alpha$. Таким образом, $N(\alpha, a) = C^s_{m^a_\alpha} C^{\ell_\alpha - s}_{L_\alpha - m^a_\alpha} / C^\ell_L$. Подставим это выражение в (11) и выделим в нём формулу гипергеометрической функции вероятности:

$$Q_\varepsilon(A) = \sum_{\alpha\in\mathfrak{A}} \sum_{a\in\alpha} \frac{1}{|\alpha|} \frac{C^{\ell_\alpha}_{L_\alpha}}{C^\ell_L} \sum_{s=s_0}^{\ell_\alpha} \big[s \le s^a_\alpha(\varepsilon)\big] \frac{C^s_{m^a_\alpha} C^{\ell_\alpha - s}_{L_\alpha - m^a_\alpha}}{C^{\ell_\alpha}_{L_\alpha}} =$$

$$= \sum_{a\in A} \sum_{\alpha\in\mathfrak{A}} \frac{[a\in\alpha]}{|\alpha|} \frac{C^{\ell_\alpha}_{L_\alpha}}{C^\ell_L} H^{\ell_\alpha, m^a_\alpha}_{L_\alpha}\big(s^a_\alpha(\varepsilon)\big).$$

Теорема доказана. ■

Следствие 2. *Пусть во множестве A найдётся алгоритм a_0, такой, что для любого $a \in A$ вектор ошибок алгоритма a_0 содержится в векторе ошибок алгоритма a. Обозначим через X_0 множество объектов, на которых ошибается алгоритм a_0. Пусть система порождающих и запрещающих множеств такова, что для всех $\alpha \in \mathfrak{A}$ выполнено $X_0 \cap X_\alpha = \varnothing$ и $X_0 \cap X'_\alpha = \varnothing$. Тогда*

$$m^a_\alpha = n(a_0, \mathbb{X}), \quad s^a_\alpha(\varepsilon) = \tfrac{\ell}{L}\big(n(a, \mathbb{X}) - \varepsilon k\big).$$

Доказательство. Зафиксируем обучающую выборку $X \in [\mathbb{X}]^\ell$ и пусть $\alpha = A(X)$. Докажем, что из $a \in \alpha$ следует $n(a, X_\alpha) = 0$. Пусть a ошибается на объекте x. Нам необходимо доказать, что $x \notin X_\alpha$. Допустим обратное, тогда по определению запрещающих объектов $x \in X_\alpha$ обязан лежать в обучении. Условие $X_0 \cap X_\alpha = \varnothing$ означает, что a_0 не ошибается на x. Следовательно, алгоритм a делает как минимум на одну ошибку больше, чем a_0 на обучении. Противоречие.

Второе утверждение заключается в том, что из $a \in \alpha$ следует $n(a, \mathbb{X}) = n(a_0, X_0) + n(a, X'_\alpha)$. Запишем число ошибок алгоритма a в виде $n(a, \mathbb{X}) = n(a, X_0) + n(a, X \setminus X_0)$. Из определения доминирующего алгоритма следует, что $n(a, X_0) = n(a_0, X_0)$. Осталось доказать, что $n(a, X \setminus X_0) = n(a, X'_\alpha)$. Отметим, что из условия $X_0 \cap X'_\alpha$ следует, что $X'_\alpha \subset X \setminus X_0$, а значит $n(a, X \setminus X_0) \geq n(a, X'_\alpha)$. Осталось доказать, что каждая ошибка $a \in X \setminus X_0$ алгоритма a принадлежит X'_α. Это следует из того, что алгоритмы a_0 и a обязаны быть неразличимыми на обучении.

Из доказанных выше утверждений следует, что

$$m^a_\alpha = n(a, \mathbb{X} \setminus X_\alpha \setminus X'_\alpha) = n(a_0, X_0) = n(a_0, \mathbb{X});$$
$$s^a_\alpha(\varepsilon) = \frac{\ell}{L}\big(n(a, \mathbb{X}) - \varepsilon k\big) - n(a, X_\alpha) = \frac{\ell}{L}\big(n(a, \mathbb{X}) - \varepsilon k\big).$$

■

Следствие 3. *Полученная формула легко объединяется с теоремой о разбиении множества алгоритмов на орбиты:*

$$Q_\varepsilon(A) = \sum_{\omega\in\Omega(A)} \sum_{\alpha\in\mathfrak{A}} [a_\omega \in \alpha] \frac{|\omega|}{|\alpha|} \frac{C_{L_\alpha}^{\ell_\alpha}}{C_L^\ell} H_{L_\alpha}^{\ell_\alpha, m_\alpha^{a_\omega}}\left(s_\alpha^{a_\omega}(\varepsilon)\right). \qquad (12)$$

Доказательство. Доказательство немедленно следует из леммы 2 о равном вкладе алгоритмов одной орбиты в вероятность переобучения. ■

2. Точные оценки вероятности переобучения

2.1. Монотонная цепочка алгоритмов

В следующих параграфах изучаются несколько модельных параметрических семейств алгоритмов, в которых количество ошибок монотонно возрастает по мере удаления вектора параметров от оптимального значения.

Определение 1. *Множество алгоритмов* $\{a_0, \ldots, a_D\}$ *называется* монотонной цепочкой*, если выполнены два условия:*

1) *монотонность числа ошибок:* $n(a_i, \mathbb{X}) = m + i,\ i = 0, \ldots, D$, *при некотором фиксированном* m*;*

2) *поглощение ошибок предыдущего алгоритма:* $\rho(a_i, a_{i-1}) = 1$, $i = 1, \ldots, D$, *где* $\rho(a, a')$ *— расстояние Хэмминга между векторами ошибок* a *и* a'*.*

Таким образом, в монотонной цепочке каждый следующий алгоритм ошибается на тех же объектах, что и предыдущий, и допускает еще одну дополнительную ошибку.

Монотонная цепочка алгоритмов — это простейшая модель однопараметрического *связного семейства алгоритмов*, предполагающая, что при непрерывном удалении некоторого параметра

от оптимального значения число ошибок на полной выборке только увеличивается.

Пример 2. *Пусть A — семейство* линейных алгоритмов классификации — *параметрических отображений из $\mathbb{X} = \mathbb{R}^n$ в $\mathbb{Y} = \{-1, +1\}$ вида*

$$a(x, w) = \operatorname{sign}(x_1 w_1 + \ldots + x_n w_n), \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

где параметр $w \in \mathbb{R}^n$ — направляющий вектор гиперплоскости, разделяющей пространство $\mathbb{R}^n$ на два полупространства — классы -1 и $+1$. Пусть функция потерь имеет вид $I(a, x) = [a(x, w) \neq y(x)]$, где $y(x)$ — истинная классификация объекта x, и множество объектов $\mathbb{X}$ линейно разделимо, т. е. существует вектор $w^* \in \mathbb{R}^n$, при котором алгоритм $a(x, w^*)$ не допускает ошибок на $\mathbb{X}$. Тогда множество алгоритмов

$$A[\delta] = \{a(x, w^* + t\delta) \colon t \in [0, +\infty)\}$$

порождает монотонную цепь при любом $\delta \in \mathbb{R}^n$, за исключением, быть может, некоторого конечного множества векторов. При этом $m = 0$ в силу линейной разделимости.

Теорема 7. *Для монотонной цепочки из $D + 1$ алгоритмов вероятность переобучения РМЭР равна*

$$Q_\varepsilon(A) = \sum_{d=0}^{D} \sum_{t=d}^{D} \frac{1}{1+t} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell', m}\big(s(\varepsilon)\big), \tag{13}$$

где $L' = L - t - F$, $\ell' = \ell - F$, $F = [t \neq D]$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m + d - \varepsilon k) \right\rfloor$.

Доказательство. в данном случае доказательство практически полностью совпадает с доказательством формулы о вероятности переобучения пессимистического ММЭР для монотонной цепи.

Действительно, перенумеруем объекты так, как показано в следующей таблице:

$$
\begin{array}{rccccccccc}
 & x_1 & x_2 & x_3 & & x_D & & \overbrace{\qquad}^{m} & \\
\boldsymbol{a}_0 = (& 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\boldsymbol{a}_1 = (& 1, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\boldsymbol{a}_2 = (& 1, & 1, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\boldsymbol{a}_3 = (& 1, & 1, & 1, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\ldots & & & & \ldots & & \ldots & \ldots & \\
\boldsymbol{a}_D = (& 1, & 1, & 1, & \ldots & 1, & 0,\ldots,0, & 1,\ldots,1 &);
\end{array}
$$

При такой нумерации каждый из алгоритмов a_t, $t = 1,\ldots,D$, допускает ошибку на объектах $x_1,\ldots,x_t$. Очевидно, лучший алгоритм a_0 не ошибается ни на одном из этих объектов. Нумерация остальных объектов не имеет значения, так как алгоритмы не различимы на них.

Заметим, что для любой обучающей выборки X множество $A(X)$ будет состоять из алгоритмов $\alpha_t \equiv \{a_0,\ldots,a_t\}$ для некоторого t. Отметим, что $|\alpha_t| = t+1$. Зафиксируем t и рассмотрим два случая:

1. Если $t = D$, то α_t совпадает со всем множеством алгоритмов, следовательно $A(X) = \alpha_t$ тогда и только тогда, когда все объекты $\{x_1,\ldots,x_D\}$ будут находиться в контрольной подвыборке $\bar{X}$. В этом случае

$$[A(X){=}\alpha_t] = [x_1,\ldots,x_D \in \bar{X}].$$

2. Во всех остальных случаях $A(X) = \alpha_t$ тогда и только тогда, когда все объекты $\{x_1,\ldots,x_t\}$ будут находиться в контрольной подвыборке $\bar{X}$, а объект x_{t+1} — в обучающей подвыборке X. В этом случае

$$[A(X){=}\alpha_t] = [x_{t+1} \in X][x_1,\ldots,x_t \in \bar{X}].$$

Для данной системы порождающих и запрещающих множеств можно применить следствие 2 из теоремы 6.

Множеству α_t и алгоритму a_d $(d \le t)$ соответствуют следующие значения параметров (для упрощения обозначений вместо двойных индексов L_{α_t} будем использовать одинарные L_t): $L_t = L - t - 1$, $\ell_t = \ell - [t = D]$, $m_t^d = m + d - d = m$, $s_t^d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$. Подставляя эти значения в общую формулу из теоремы 6 о порождающих и запрещающих объектах, получаем утверждение нашей теоремы. ■

В приведенном доказательстве мы не рассматривали отдельно случай $D \le k$ и $D > k$. Эти эффекты уже учтены корректно благодаря доопределению нулем биномиальных коэффициентов и гипергеометрического распределения. Так же отметим, что в случае монотонной цепочки группа симметрии в данном случае была тривиальной, и потому не учитывалась при вычисления.

2.2. Унимодальная цепочка алгоритмов

Определение 2. *Множество алгоритмов*

$$\{a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_D\}$$

называется унимодальной цепочкой, если выполнены два условия:

1) *левая ветвь* $\{a_0, a_1, \ldots, a_D\}$ *и правая ветвь* $\{a_0, a'_1, \ldots, a'_D\}$ *являются монотонными цепочками.*

2) *пересечение множества ошибок алгоритмов* a_D *и* a'_D *равно множеству ошибок алгоритма* a_0.

Параметр D будем называть *длиной ветвей* унимодальной цепочки.

Унимодальная цепочка является более реалистичной моделью однопараметрического *связного семейства* по сравнению с монотонной цепочкой. Если мы имеем лучший алгоритм a_0 с оптимальным значением некоторого вещественного параметра, то отклонение значения этого параметра как в бóльшую, так и в меньшую стороны приводит к увеличению числа ошибок.

Теорема 8. *Для унимодальной цепочки с ветвями длины D вероятность переобучения рандомизированного метода минимизации эмпирического риска равна*

$$Q_\varepsilon(A) = \sum_{d=0}^{D} \sum_{t_1=d}^{D} \sum_{t_2=0}^{D} \frac{|\omega_d|}{1+t_1+t_2} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}\big(s(\varepsilon)\big), \tag{14}$$

где $\omega_d = [d=0] + 2\cdot[d>0]$, $L' = L - S - F$, $S = t_1 + t_2$, $F = [t_1 \neq D] + [t_2 \neq D]$, $\ell' = \ell - F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m+d-\varepsilon k) \right\rfloor$. *[?].*

Доказательство.

Пронумеруем объекты генеральной выборки $\mathbb{X}$ таким образом, как показано в следующей таблице:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \\ x_D \\ \hline x'_1 \\ x'_2 \\ \\ x'_D \end{array} \begin{array}{c} \begin{array}{cccccccccc} a_0 & a_1 & a_2 & \cdots & a_D & a'_1 & a'_2 & \cdots & a'_D \end{array} \\ \left(\begin{array}{cccccccccc} 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right) \end{array}$$

Нумерация остальных объектов не имеет значения, так как алгоритмы не различимы на них.

Перестановками объектов выборки $(x_1 \leftrightarrow x'_1, \ldots, x_D \leftrightarrow x'_D)$ можно поменять левую и правую ветви местами. Следовательно, алгоритмы разных ветвей с равным числом ошибок лежат в одной орбите действия группы симметрии. Орбита $\omega_0 = \{a_0\}$ содержит единственный алгоритм. Для остальных орбит $\omega_d = \{a_d, a'_d\}$ договоримся выбирать алгоритм a_d из левой ветви в качестве представителя орбиты.

Для произвольной обучающей выборки $X \in [\mathbb{X}]^\ell$ множество $A(X)$ будет состоять из алгоритмов $a_0, a_1, \ldots, a_{t_1}, a'_1, \ldots, a'_{t_2}$

для некоторой пары (t_1, t_2). Следовательно, $\mathfrak{A} = \{\alpha_{t_1,t_2}\}$, где $t_i = 0, \ldots, D$. Отметим, что $|\alpha_{t_1,t_2}| = \frac{1}{1+t_1+t_2}$, а $[a_d \in \alpha_{t_1,t_2}] = [d \le t_1]$.

Для $\alpha \equiv \alpha_{t_1,t_2}$ при t_1, t_2 строго меньше D, множеством порождающих объектов будет $X'_\alpha = \{x_{t_1+1}, x_{t_2+1}\}$. Условие $t_i = D$ уменьшает количество порождающих объектов в X'_α на единицу. Множество запрещающих объектов $X_\alpha = \{x_1, \ldots, x_{t_1}, x'_1, \ldots, x'_{t_2}\}$.

Введя обозначение $F = [t_1 \neq D] + [t_2 \neq D]$, получим

$$L_\alpha = L - t_1 - t_2 - F, \qquad \ell_\alpha = \ell - F;$$

$$m^a_\alpha = m + d - d = m, \quad s^a_\alpha(\varepsilon) = \lfloor \frac{\ell}{L}(m + d - \varepsilon k) \rfloor.$$

Подставляя эти значения в общую формулу из теоремы 6 о порождающих и запрещающих объектах, получаем утверждение нашей теоремы.

■

2.3. Пучок монотонных цепочек

Пучком из h монотонных цепочек называется множество алгоритмов, полученное объединением h монотонных цепочек равной длины, с общим первым алгоритмом. Как и в случае унимодальной цепочки, предполагается, что множества объектов, на которых ошибаются алгоритмы ветвей, не пересекаются.

Связка из $2h$ монотонных цепочек является моделью h-параметрического семейства алгоритмов, в котором разрешено изменять любой из h параметров при фиксированных остальных, а одновременное изменение нескольких параметров не допускается. Данное семейство можно также рассматривать как обобщение трёх частных случаев, рассмотренных в [**?**]: монотонной цепочки ($h = 1$), унимодальной цепочки ($h = 2$) и единичной окрестности лучшего алгоритма ($D = 1$).

В следующей теореме будет дана явная формула вероятности переобучения для связки из h монотонных цепочек. Введём

комбинаторный коэффициент $R^d_{D,h}(S,F)$, который зависит от параметров S и F, от числа монотонных цепочек h и от их длины D, а также от d — минимального значения параметра S. Коэффициент $R^d_{D,h}(S,F)$ равен числу способов представить число S в виде суммы h неотрицательных пронумерованных слагаемых, $S = t_1 + \ldots + t_h$, каждое из которых не превосходит D. При этом ровно F слагаемых не должно равняться D, а на первое слагаемое накладывается дополнительное ограничение $t_1 \geq d$.

Теорема 9. *Пусть в связке из h монотонных цепочек лучший алгоритм допускает m ошибок на полной выборке, длина каждой ветви без учета лучшего алгоритма равна D. Тогда при обучении рандомизированным методом вероятность переобучения может быть записана в виде:*

$$Q_\varepsilon(A) = \sum_{d=0}^{D} \sum_{S=d}^{hD} \sum_{F=0}^{h} \frac{|\omega_d| R^d_{D,h}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}\big(s(\varepsilon)\big), \qquad (15)$$

где $L' = L - S - F$, $\ell' = \ell - F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m + d - \varepsilon k) \right\rfloor$; $|\omega_h| = 1$ *при* $h = 0$ *и* $|\omega_d| = h$ *при* $d \geq 1$.

Доказательство.

Группа симметрии связки из h монотонных цепочек является симметрической группой S_h, действующей на ветви связки всевозможными перестановками. Следовательно, алгоритмы разных ветвей с равным числом ошибок лежат в одной орбите действия группы симметрии. Орбита $\omega_0 = \{a_0\}$ содержит единственный алгоритм. Для остальных орбит $\omega_d = \{a^1_d, a^2_d, \ldots, a^h_d\}$ договоримся выбирать алгоритм a^1_d из первой ветви в качестве представителя орбиты.

Для произвольной обучающей выборки $X \in [\mathbb{X}]^\ell$ множество $A(X)$ будет состоять из алгоритмов $\{a^i_j\}$ при $i = 1, \ldots, h$, $j = 1, \ldots, t_i$, для некоторого вектора $\boldsymbol{t} \equiv (t_1, t_2, \ldots, t_h)$, где $t_i = = 0, \ldots, D$. Отметим, что $|\alpha_{\boldsymbol{t}}| = \frac{1}{1+t_1+\ldots+t_h}$, а $[a^1_d \in \alpha_{\boldsymbol{t}}] = [d \leq t_1]$.

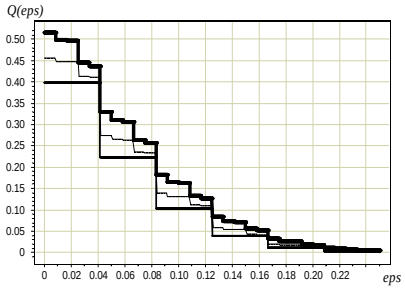


Рис. 1. Зависимость $Q_\varepsilon(A)$ от ε для монотонной цепочки при $L = 100$, $\ell = 60$, $D = 40$, $m = 20$.

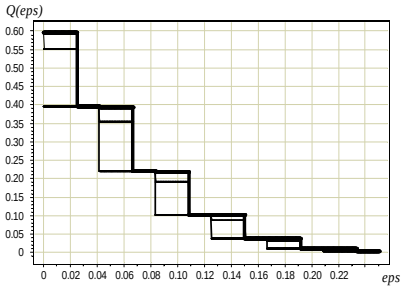


Рис. 2. Зависимость $Q_\varepsilon(A)$ от ε для единичной окрестности при $L = 100$, $\ell = 60$, $h = 10$, $m = 20$.

Для $\alpha_{\boldsymbol{t}}$, где все t_i строго меньше D, множеством порождающих объектов будет $X'_\alpha = \{x^1_{t_1+1}, \ldots, x^h_{t_h+1}\}$. Условие $t_i = D$ уменьшает количество порождающих объектов в X'_α на единицу. Множество запрещающих объектов $X_\alpha = \{x^1_1, \ldots, x^1_{t_1}, x^2_1, \ldots, x^2_{t_2}, \ldots, x^h_1, \ldots, x^h_{t_h}\}$.

Введем обозначения $S = \sum\limits_{i=1}^{p} t_i$, $F = \sum\limits_{i=1}^{p} [t_i \neq D]$. Тогда $L_\alpha = L - S - F$, $\ell_\alpha = \ell - F$, $m^a_\alpha = n(a_0, \mathbb{X})$, $s^a_\alpha = \lfloor \frac{\ell}{L}(m + d - \varepsilon k) \rfloor$.

Подставляя эти значения в общую формулу из теоремы 6 о порождающих и запрещающих объектах, получаем следующее выражение для вероятности переобучения:

$$Q_\varepsilon(A) = \sum_{d=0}^{D} \sum_{t_1=d}^{D} \sum_{t_2=0}^{D} \ldots \sum_{t_h=0}^{D} \frac{|\omega_d|}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}\big(s(\varepsilon)\big).$$

Теперь от суммирования по параметрам t_i можно перейти

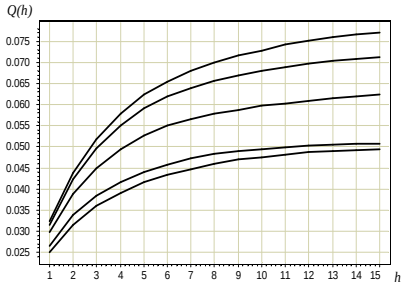


Рис. 3. Зависимость $Q_\varepsilon(A)$ от h для связки из монотонных цепочек при $L = 300$, $\ell = 150$, $m = 15$, $D = 1, 2, 3, 5, 10$, $\varepsilon = 0.05$.

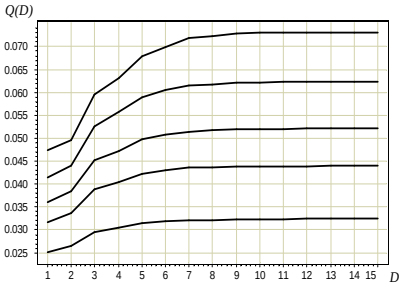


Рис. 4. Зависимость $Q_\varepsilon(A)$ от D для связки из $h = 1, 2, 3, 5, 10$ монотонных цепочек при $L = 300$, $\ell = 150$, $m = 15$, $\varepsilon = 0.05$.

к суммированию по множеству возможных значений S и F:

$$Q_\varepsilon(A) = \sum_{d=0}^{D} \sum_{S=d}^{hD} \sum_{F=0}^{h} |\omega_h| \frac{R^d_{D,h}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}\big(s(\varepsilon)\big),$$

где $R^d_{D,h}(S,F)$ — определенный выше комбинаторный коэффициент.

■

Следствие 4. *Для единичной окрестности из h алгоритмов вероятность переобучения равна*

$$Q_\varepsilon(A) = \sum_{d=0}^{1} \sum_{S=d}^{h} \frac{|\omega_d| C^{S-d}_{h-d}}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}\big(s(\varepsilon)\big), \tag{16}$$

где $L' = L-h$, $\ell' = \ell+S-h$.

На рис. 1 и рис. 2 представлены результаты численных экспериментов, в которых сравнивались вероятности переобучения

для различных вариантов минимизации эмпирического риска. Из четырех кривых на каждом графике верхняя (жирная) соответствует пессимистической минимизации эмпирического риска [**?**, **?**], нижняя — оптимистической. Две почти сливающиеся кривые между ними соответствуют рандомизированной минимизации эмпирического риска. Одна из них вычислена по доказанным формулам, вторая построена методом Монте-Карло по 10^5 случайных разбиений, при равновероятном выборе лучшего алгоритма в случаях неопределенности. Различия этих двух кривых находятся в пределах погрешности метода Монте-Карло.

На рис. 3 и рис. 4 представлены зависимости вероятности переобучения от числа h ветвей в связке и от их длины D. Графики построены для рандомизированного метода минимизации эмпирического риска. Рис. 4 показывает, что при увеличении длин цепочек D вероятность переобучения практически перестаёт расти уже при $D = 7$. Это связано с *эффектом расслоения* — лишь алгоритмы из нижних слоёв имеют существенно отличную от нуля вероятность быть выбранными методом минимизации эмпирического риска. Добавление «слишком плохих» алгоритмов не увеличивает вероятность переобучения. Рис. 3 показывает, что при увеличении числа h цепочек в связке вероятность переобучения продолжает расти. Однако скорость роста сублинейна по h, благодаря *эффекту связности* — все алгоритмы находятся на хэмминговом расстоянии не более D от лучшего алгоритма.

2.4. Многомерная монотонная сеть алгоритмов

Введём целочисленный вектор индексов $\boldsymbol{d} = (d_1, \ldots, d_h) \in \mathbb{Z}^h$. Обозначим $\|\boldsymbol{d}\| = \max\limits_{j=1,\ldots,h} |d_j|$, $|\boldsymbol{d}| = |d_1| + \ldots + |d_h|$. На множестве векторов индексов введём покомпонентное отношение сравнения: $\boldsymbol{d} < \boldsymbol{d}'$, если $d_j \leq d'_j$, $j = 1, \ldots, h$, и хотя бы одно из неравенств строгое.

Определение 3. *Множество алгоритмов* $A = \{a_{\boldsymbol{d}}\}$, *где* $\boldsymbol{d} \geq 0$

и $\|\boldsymbol{d}\| \leq D$ *называется* монотонной h-мерной сеткой алгоритмов длины D, *если существует* $h \in \mathbb{N}$ *и упорядоченные наборы объектов* $X_j = \{x_j^1, \ldots, x_j^D\} \subset \mathbb{X}$, *для всех* $j = 1, \ldots, h$, *а так же множества* $U_1 \subset \mathbb{X}$ *и* $U_0 \subset \mathbb{X}$, *такие что:*

1) *набор* $\left\{U_0, U_1, \{X_j\}_{j=1}^h\right\}$ *является разбиением множества* $\mathbb{X}$ *на непересекающиеся подмножества;*

2) $a_d(x_j^i) = [i \leq d_j]$, *где* $x_j^i \in X_j$;

3) $a_d(x_0) = 0$ *при всех* $x_0 \in U_0$;

4) $a_d(x_1) = 1$ *при всех* $x_1 \in U_1$.

Монотонная сетка алгоритмов — это модель параметрического *связного семейства алгоритмов*, предполагающая, что при непрерывном удалении каждой компоненты вектора параметров от оптимального значения число ошибок на полной выборке только увеличивается.

Обозначим $|U_1| = m$. Из определения следует, что $n(a_{\boldsymbol{d}}, \mathbb{X}) = m + |d|$. Алгоритм $a_{\mathbf{0}}$ является *лучшим в сетке*. Множество алгоритмов с равным числом ошибок $t + m = n(a_{\boldsymbol{d}}, \mathbb{X})$ называются t-*слоем* сетки.

Пример 3. *Монотонная двумерная сетка при* $m = 0$ *и* $L = 4$:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccccccccc} a_{0,0} & a_{1,0} & a_{2,0} & a_{0,1} & a_{1,1} & a_{2,1} & a_{0,2} & a_{1,2} & a_{2,2} \end{array} \\ \left(\begin{array}{ccccccccc} 0 & \boldsymbol{1} & \boldsymbol{1} & 0 & \boldsymbol{1} & \boldsymbol{1} & 0 & \boldsymbol{1} & \boldsymbol{1} \\ 0 & 0 & \boldsymbol{1} & 0 & 0 & \boldsymbol{1} & 0 & 0 & \boldsymbol{1} \\ \hline 0 & 0 & 0 & \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{1} \end{array}\right) \end{array}$$

Число алгоритмов в h-мерной монотонной сетке с ветвями длины D равно $(D+1)^h$. *Укороченной* h-мерной монотонной сеткой $\tilde{A} \subset A$ назовем первые D слоев из A. Таким образом

$$\tilde{A} = \{a_{\boldsymbol{d}} \in A,\ |\boldsymbol{d}| \leq D\}.$$

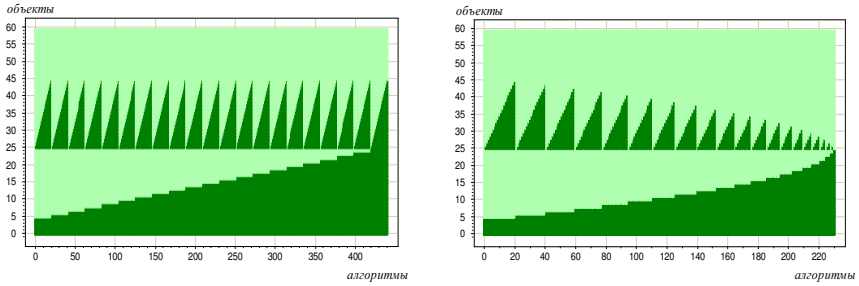


Рис. 5. Матрица ошибок монотонной сетки (слева) и укороченной монотонной сетки (справа) при $D = 20$, $h = 2$, $m = 5$, $L = 60$.

Число алгоритмов в $\tilde{A}$ равно C^h_{D+h}.

Впервые монотонны сетки произвольной размерности были изучены П. Ботовым в [**?**]. Там же были получены формулы для вероятности переобучения *пессимистического* метода минимизации эмпирического риска.

Численные эксперименты показывают, что при разумных сочетаниях параметров вероятности переобучения для укороченной $\tilde{A}$ и простой A монотонных сеток различаются крайне мало. Поэтому в дальнейшем мы ограничимся исследованием неукороченных монотонных сеток. Для этого класса семейств алгоритмов будут получены явные формулы вероятности переобучения рандомизированного метода минимизации эмпирического риска.

Теорема 10. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к монотонной сети $A = \{a_{\boldsymbol{d}}\}$ размерности h, $\|\boldsymbol{d}\| \le D$, дается выражением:*

$$Q_\varepsilon(A) = \sum_{\substack{\boldsymbol{d}\ge\boldsymbol{0},\\ \|\boldsymbol{d}\|\le D}} \sum_{\substack{\boldsymbol{t}\ge\boldsymbol{0},\\ \|\boldsymbol{t}\|\le D}} \frac{[\boldsymbol{t}\ge\boldsymbol{d}]}{V(\boldsymbol{t})} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}\big(s(\varepsilon)\big), \tag{17}$$

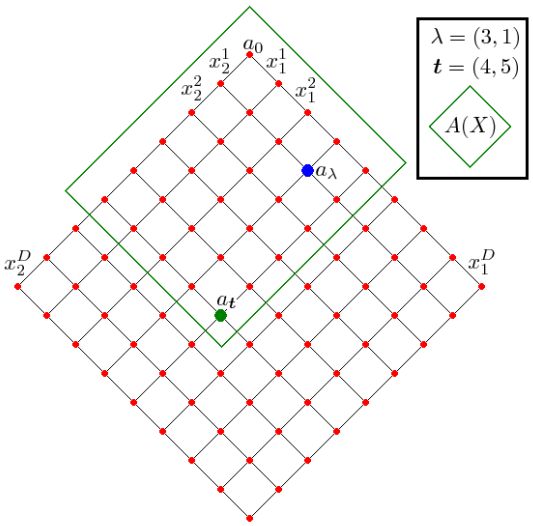


Рис. 6. Строение множества $A(X)$ для двумерной монотонной сети; $h = 2$, $D = 8$.

где $V(\boldsymbol{t}) = \prod\limits_{j=1}^{h}(t_j+1)$, $\ell' = \ell - \sum\limits_{j=1}^{h}[t_j \neq D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$, $s(\varepsilon) = \frac{\ell}{L}\big[m + |\boldsymbol{d}| - \varepsilon k\big]$.

Доказательство. Напомним, что через $\mathfrak{A} = \big\{A(X)\colon X \in [\mathbb{X}]^\ell\big\}$ обозначалось множество подмножеств алгоритмов, получающихся в результате обучения. Пусть A — монотонная сеть. Тогда для произвольной обучающей выборки X множество $A(X)$ устроено специфическим образом. На рис. 6 показано, что в $A(X)$ всегда найдется такой алгоритм $a_{\boldsymbol{t}}$, что $A(X) = \{a_{\boldsymbol{d}} \,|\, \boldsymbol{d} \leq \boldsymbol{t}\}$. Следовательно, $\mathfrak{A} = \{\alpha_{\boldsymbol{t}} : \boldsymbol{t} \in [0, \ldots, D]^h\}$.

Обозначим через $J(\boldsymbol{t})$ — множество тех индексов $j \in \{1, \ldots, h\}$,

для которых $t_j < D$. Положим

$$\bar{X}_{\boldsymbol{t}} = \bigcup_{j \in J(\boldsymbol{t})} x_j^{t_j+1}, \quad \bar{X}'_{\boldsymbol{t}} = \bigcup_{j=1}^{h} \bigcup_{i=1}^{t_j} x_j^i.$$

Построенные таким образом $\bar{X}_{\boldsymbol{t}}$ и $\bar{X}'_{\boldsymbol{t}}$ являются порождающим и запрещающим множеством для $\alpha_{\boldsymbol{t}}$. Применив теорему о порождающих и запрещающих множествах, получим формулу (17). ■

Вычисление вероятности переобучения по полученной формуле требует $O(h \cdot D^{2h})$ операций. Теорема 12 позволяет сократить количество вычислений за счет учета симметрий монотонной сети.

Лемма 4. *Группа симметрии монотонной сетки размерности h содержит в качестве подгруппы группу S_h всевозможных перестановок множеств $X_1, \ldots, X_h$.*

Доказательство.

Все алгоритмы h-мерной монотонной сетки длины D индексированы множеством вектор-индексов $\boldsymbol{d} \in \{0, \ldots, D\}^h$. Тут число ошибок алгоритма $a_{\boldsymbol{d}} = m + |\boldsymbol{d}|$.

Рассмотрим алгоритм $a_{\boldsymbol{d}} \in A$ и произвольную $\pi \in S_h$. По данному выше определению действия π на $\mathbb{X}$ получаем, что $\pi a_{\boldsymbol{d}} = a_{\pi \boldsymbol{d}}$, где действие π на вектор $\boldsymbol{d}$ определяется соответствующей перестановкой его координат. Множество $\{0, \ldots, D\}^h$ сохраняется при применении к нему произвольной перестановки координат $\pi \in S_h$. Поэтому $\forall \boldsymbol{d} \in \{0, \ldots, D\}^h$ выполнено $\pi \boldsymbol{d} \in \{0, \ldots, D\}^h$. А следовательно, $a_{\pi \boldsymbol{d}} \in A$. ■

Пусть Y_h^D — множество целочисленных неотрицательных невозрастающих последовательностей длины h и не превосходящих D, $|S_h \boldsymbol{d}|$ — число различных слов, состоящих из символов $d_1, \ldots, d_h$.

Лемма 5. *Множество орбит монотонной сетки $A = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leq D$ под действием S_h индексировано всевозможными векторами $\lambda \in Y_h^D$. Число алгоритмов в орбите ω_λ, где $\lambda = (\lambda_1, \ldots, \lambda_h)$ равно числу различных слов длины h, состоящих из символов $\lambda_1, \ldots, \lambda_h$: $|\omega_\lambda| = |S_h\lambda|$.*

Доказательство. Напомним, что вместо действия S_h на $A = \{a_{\boldsymbol{d}}\}$ можно рассматривать действие S_h на вектор индексов $\boldsymbol{d}$, заданное перестановками координат.

Рассмотрим орбиту произвольного алгоритма $a_{\boldsymbol{d}}$. Возьмем перестановку $\pi \in S_h$, упорядочивающую координаты $\boldsymbol{d}$ в порядке не-возрастания, и положим $\lambda = \pi\boldsymbol{d}$. Построенная таким образом λ лежит в множестве Y_h^D. При этом различным λ_1 и λ_2 будут соответствовать различные орбиты действия группы S_h на $\{a_{\boldsymbol{d}}\}$.

Взаимно-однозначное соответствие между словами длины h из символов $\lambda_1, \ldots, \lambda_h$ и количеством элементами орбиты $|\omega_\lambda|$ очевидно. ■

Теорема 11. *С учетом симметрий монотонной сети вероятность переобучения записывается в виде*

$$Q_\varepsilon(A) = \sum_{\boldsymbol{d} \in Y_h^D} \sum_{\substack{\boldsymbol{t} \geq \boldsymbol{d}, \\ \|\boldsymbol{t}\| \leq D}} \frac{|S_h\boldsymbol{d}|}{V(\boldsymbol{t})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}\big(s(\varepsilon)\big), \tag{18}$$

где Y_h^D — множество целочисленных неотрицательных невозрастающих последовательностей длины h и не превосходящих D, $|S_h\boldsymbol{d}|$ — число различных слов, состоящих из символов $d_1, \ldots, d_h$.

Доказательство. Для доказательства формулы (18) необходимо воспользоваться теоремой 12 и леммой 5. ■

Расчет новой формулы требует $O(h \cdot D^h \cdot C_{D+h}^h)$ операций. Рассмотрим отношение D^h / C_{h+D}^h, показывающее во сколько раз сокращается объем вычислений благодаря учету симметрии. Данная величина максимальна при $D \gg h$. Это соответствует случаю

сеток большой длины, на которых группа симметрии действует наиболее эффективно. В этом случае число операций сокращается в $h!$ раз, что в точности соответствует количеству элементов в группе симметрий. В остальных случаях (сетки больших размерностей и малой длины) выигрыш оказывается меньше.

2.5. Многомерная унимодальная сеть алгоритмов

Унимодальная сетка является более реалистичной моделью связного параметрического семейства по сравнению с монотонной сеткой. Если мы имеем лучший алгоритм a_0 с оптимальным значением вектора вещественных параметров, то отклонение значений компонент этого вектора как в бóльшую, так и в меньшую сторону приводит к увеличению числа ошибок.

Определение 4. *Множество алгоритмов* $A = \{a_{\boldsymbol{d}}\}$, *где* $\|\boldsymbol{d}\| \le D$ *называется* унимодальной h-мерной сеткой алгоритмов, *если существует* $h \in \mathbb{N}$ *и упорядоченные наборы объектов* $X_j = \{x_j^1, x_j^2, \ldots, x_j^D\} \subset \mathbb{X}$, $Y_j = \{y_j^1, y_j^2, \ldots, y_j^D\} \subset \mathbb{X}$, *для всех* $j = 1, \ldots, h$, *а также множества* $U_1 \subset \mathbb{X}$ *и* $U_0 \subset \mathbb{X}$, *такие что выполнены условия:*

1) *Набор* $\{U_0, U_1, \{X_j\}_{j=1}^h, \{Y_j\}_{j=1}^h\}$ *является разбиением множества* $\mathbb{X}$ *на непересекающиеся множества;*

2) $a_d(x_j^i) = [d_j > 0]\big[i \le |d_j|\big]$, *где* $x_j^i \in X_j$;

3) $a_d(y_j^i) = [d_j < 0]\big[i \le |d_j|\big]$, *где* $y_j^i \in Y_j$;

4) $a_d(x_0) = 0$ *при всех* $x_0 \in U_0$;

5) $a_d(x_1) = 1$ *при всех* $x_1 \in U_1$.

Заметим, что данное определение отличается от определения монотонной сетки отсутствием ограничения $\boldsymbol{d} \ge 0$. Число алгоритмов в h-мерной унимодальной сетке с ветвями длины D составляет $(2D+1)^h$. *Укороченной* h-мерной унимодальной сеткой

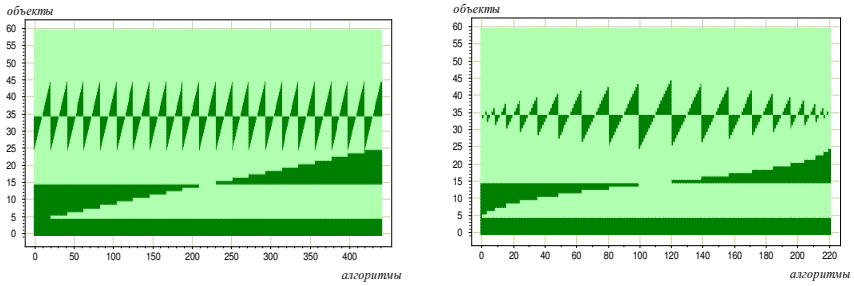


Рис. 7. Матрица ошибок унимодальной сетки (слева) и укороченной унимодальной сетки (справа) при $D = 10$, $h = 2$, $m = 5$, $L = 60$.

$\tilde{A}$ назовем множество первых D слоев из A:

$$\tilde{A} = \{a_{\boldsymbol{d}} \in A \colon n(a_{\boldsymbol{d}}, \mathbb{X}) \leq m + D\}.$$

Формула для вероятности переобучения *пессимистического* метода минимизации эмпирического риска на укороченных унимодальных сетках так же была получена в [**?**]. Ниже рассматриваются неукороченные унимодальные сетки и случай *рандомизированного* метода минимизации эмпирического риска.

Лемма 6. *Группа симметрии унимодальной сетки размерности h содержит в качестве подгруппы группу* $\mathrm{Sym}(A) = (S_2)^h \times S_h$. *Группа S_h действует на множестве пар $(X_j, Y_j)_{j=1}^h$ всеми возможными перестановками; j-тая группа S_2 переставляет объекты множества X_j и Y_j местами, сохраняя относительный порядок объектов.*

Доказательство. Все алгоритмы h-мерной унимодальной сетки длины D индексированы множеством вектор-индексов $\boldsymbol{d} \in \in \{-D, \ldots, D\}^h$. Тут число ошибок алгоритма $a_{\boldsymbol{d}} = m + |\boldsymbol{d}|$.

Рассмотрим алгоритм $a_{\boldsymbol{d}} \in A$ и произвольную $\pi = = (z_1, \ldots, z_h) \times \pi_0 \in \mathrm{Sym}(A)$, где $z_j \in S_2$, $\pi_0 \in S_h$. По данному

выше определению действия π на $\mathbb{X}$ получаем, что $\pi a_{\boldsymbol{d}} = a_{\pi \boldsymbol{d}}$, где действие π на вектор $\boldsymbol{d}$ определяется перестановкой его координат с помощью π_0 и инверсией знаков для всех j, таких что $z_j \neq id$ — транспозиция. Множество $\{-D, \ldots, D\}^h$ сохраняется при применении к нему произвольной перестановки координат $\pi \in (S_2)^h \times S_h$. Поэтому $\forall \boldsymbol{d} \in \{-D, \ldots, D\}^h$ выполнено $\pi \boldsymbol{d} \in \{-D, \ldots, D\}^h$. А следовательно, $a_{\pi \boldsymbol{d}} \in A$. ■

Лемма 7. *Множество орбит унимодальной сетки $A = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leq D$, под действием* $\mathrm{Sym}(A)$ *индексировано всевозможными Y_h^D. Пусть $\lambda = (\lambda_1, \ldots, \lambda_h) \in Y_h^D$. Обозначим через $|S_h \lambda|$ число различных слов длины h, состоящих из символов $\lambda_1, \ldots, \lambda_h$ Пусть $|\lambda > 0|$ — число строго положительных компонент вектора λ.*

Тогда число алгоритмов в орбите ω_λ равно $|S_h \lambda| \cdot 2^{|\lambda > 0|}$.

□ **Доказательство** полностью повторяет рассуждения леммы 5. Множитель $2^{|\lambda > 0|}$ соответствует возможности сменить знак у всех не-нулевых компонент вектора $\boldsymbol{d}$. ■

Теорема 12. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к унимодальной сетке $A = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leq D$, дается выражением:*

$$Q_\varepsilon(A) = \sum_{\boldsymbol{d} \in Y_h^D} \sum_{\substack{\boldsymbol{t} \geq \boldsymbol{d}, \\ \|\boldsymbol{t}\| \leq D}} \sum_{\substack{\boldsymbol{t}' \geq 0, \\ \|\boldsymbol{t}'\| \leq D}} \frac{|S_h \boldsymbol{d}\,| \cdot 2^{|\boldsymbol{d} > 0|}}{T(\boldsymbol{t} + \boldsymbol{t}')} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell', m}(s_0), \qquad (19)$$

где $\ell' = \ell - \sum_{j=1}^h \Big([t_j \neq D] + [t'_j \neq D] \Big)$, $k' = k - |\boldsymbol{t}| - |\boldsymbol{t}'|$, а остальные обозначения совпадают с обозначениями теоремы 11.

Доказательство. Напомним, что через $\mathfrak{A} = \big\{ A(X) \colon X \in [\mathbb{X}]^\ell \big\}$ обозначалось множество подмножеств алгоритмов, получающихся в результате обучения. Пусть A — унимодальная сеть. Тогда

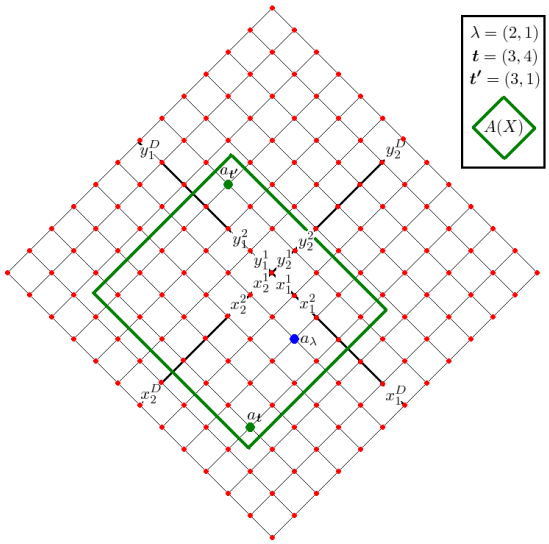


Рис. 8. Строение множества $A(X)$ для двумерной унимодальной сетки.

для произвольной обучающей выборки X множество $A(X)$ устроено специфическим образом. На рис. 8 показано, что в $A(X)$ всегда найдется такая пара алгоритмов $(a_{\boldsymbol{t}_1},\, a_{\boldsymbol{t}_2})$ что

$$A(X) = \{a_{\boldsymbol{d}} \,|\, \boldsymbol{t}_1 \leq \boldsymbol{d} \leq \boldsymbol{t}_2\}.$$

Следовательно, $\mathfrak{A} = \{\alpha_{\boldsymbol{t},\boldsymbol{t}'} : \boldsymbol{t}, \boldsymbol{t}' \in [0, \ldots, D]^h\}$.

Обозначим через $J(\boldsymbol{t})$ — множество тех индексов $j \in \{1, \ldots, h\}$, для которых $t_j < D$. Положим

$$X_{\boldsymbol{t}} = \bigcup_{j \in J(\boldsymbol{t})} x_j^{t_j+1}, \quad X'_{\boldsymbol{t}} = \bigcup_{j=1}^{h} \bigcup_{i=1}^{t_j} x_j^i;$$

$$Y_{\boldsymbol{t}'} = \bigcup_{j \in J(\boldsymbol{t}')} y_j^{t'_j+1}, \quad Y'_{\boldsymbol{t}'} = \bigcup_{j=1}^{h} \bigcup_{i=1}^{t'_j} y_j^i.$$

Множества $X_{\boldsymbol{t}} \cup Y_{\boldsymbol{t}'}$ и $X'_{\boldsymbol{t}} \cup Y'_{\boldsymbol{t}'}$ являются соответственно порождающим и запрещающим множеством для $\alpha_{\boldsymbol{t},\boldsymbol{t}'}$. Применив теорему о порождающих и запрещающих множествах, получим формулу (19). ■

2.6. Разреженные монотонные и унимодальные сети

В предыдущих параграфах рассматривались семейства алгоритмов, реализующихся при непрерывном изменении компонент вектора вещественных параметров. На практике возможны ситуации, при которых наблюдаемое семейство будет собственным подмножеством рассмотренных выше монотонных и унимодальных сетей. В данном параграфе рассматриваются только такие подмножества, в которых наложено ограничение на минимальное расстояние между ближайшими алгоритмами в семействе. Такие случаи соответствует изменению каждой компоненты вектора вещественных параметров с постоянным шагом.

Определение 5. *Пусть $\rho \in \mathbb{N}$ — целочисленный параметр; $A = \{a_{\boldsymbol{d}}\}$ — h-мерная монотонная сетка длины ρD; $m \equiv n(a_0, \mathbb{X})$. Разреженной h-мерной монотонной сеткой $\ddot{A}$ плотности ρ и длины D будем называть подмножество A, заданное условием:*

$$\ddot{A} = \left\{ a_{\boldsymbol{d}} \in A \,\middle|\, \boldsymbol{d} \in (\rho\mathbb{Z})^h \right\}.$$

Отметим, что при $\rho > 1$ граф смежности разреженной монотонной сетки состоит из изолированных точек.

Пример 4. *На рисунке 4 выделено подмножество двумерной монотонной сетки с параметром $D = 8$, соответствующее разреженной монотонной сетке с параметрами $\rho = 2$, $D = 4$.*

Определение 6. *Пусть $\rho \in \mathbb{N}$ — целочисленный параметр; $A = \{a_{\boldsymbol{d}}\}$, где $\|d\| \leq \rho D$ — h-мерная унимодальная сетка; $m \equiv$*

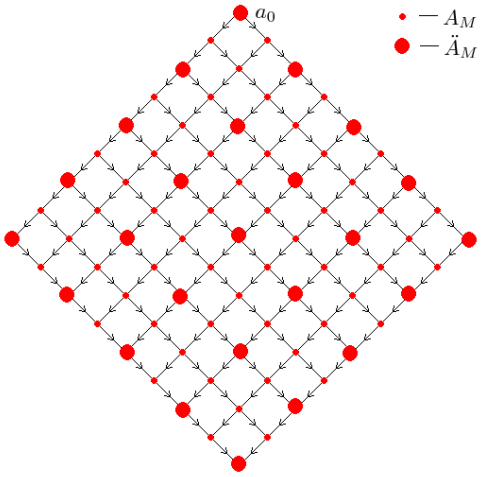


Рис. 9. Двумерная разреженная монотонная сетка при $\rho = 2$, $D = 4$.

$\equiv n(a_0, \mathbb{X})$. Разреженной h-мерной унимодальной сеткой $\ddot{A}$ плотности ρ *будем называть следующее подмножество* A*:*

$$\ddot{A} = \left\{ a_{\boldsymbol{d}} \in A \,\middle|\, \boldsymbol{d} \in (\rho\mathbb{Z})^h \right\}.$$

Теорема 13. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к разреженной монотонной сетке* $\ddot{A}_M = \{a_{\boldsymbol{d}}\}$ *размерности* h, $\|d\| \leq D$, *дается выражением:*

$$Q_\varepsilon(\ddot{A}_M) = \sum_{\lambda \in Y_*^{h,D}} \sum_{\substack{\boldsymbol{t} \geq \rho\lambda, \\ \|\boldsymbol{t}\| \leq \rho D}} \frac{|S_h\lambda|}{T\big(\lfloor \boldsymbol{t}/\rho \rfloor\big)} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}(s_0), \tag{20}$$

где Y_h^D *— множество целочисленных неотрицательных невозрастающих последовательностей длины* h *и не превосходящих* D, $|S_h\lambda|$ *— мощность орбиты действия симметрической группы* S_h

на λ, $T(\boldsymbol{t}) = \prod_j (t_j + 1)$, $\ell' = \ell - \sum_{j=1}^{h} [t_j \neq \rho D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$, $s_0 = \frac{\ell}{L}[m + \rho|\lambda| - \varepsilon k]$, $H_{L'}^{\ell',m}(s)$ — *функция гипергеометрического распределения.*

Доказательство. Воспользуемся теоремой 1 о разложении вероятность переобучения по орбитам множества алгоритмов:

$$Q_\varepsilon(A_M) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_*^{h,D}} |S_h \lambda| \sum_{X \in [\mathbb{X}]^\ell} \frac{[a_\lambda \in A_M(X)]}{|A_M(X)|} [\delta(a_\lambda, X) \geq \varepsilon] .$$

Шаг 1. Зафиксируем $X \in [\mathbb{X}]^\ell$. Обозначим через t_j максимальный индекс из $\{0, \ldots, \rho D\}$, при котором все объекты $\{x_j^1, \ldots, x_j^{t_j}\}$ содержатся в $\bar{X}$, а $x_j^{t_j+1}$, при его наличии, лежит в X. Положим $\boldsymbol{t} = \{t_j\}_{j=1}^h$. Тогда условие $a_\lambda \in A_M(X)$ перепишется как $\boldsymbol{t} \geq \rho\lambda$.

Действительно, заметим, что для всех $a \in A_M$ и $X \in [\mathbb{X}]^\ell$ выполнено $n(a, X) \geq n(a_0, X)$. Следовательно, алгоритм a_λ может быть выбран, только если объекты x_j^i при всех $j = 1, \ldots, h$ и $i \leq \rho\lambda_j$ лежат в контроле. В терминах $\boldsymbol{t}$ это записывается как $\boldsymbol{t} \geq \rho\lambda$.

Обозначим множество разбиений на обучение и контроль с фиксированным значением параметра $\boldsymbol{t}$ через $[\mathbb{X}]_{\boldsymbol{t}}^\ell$. Тогда

$$Q_\varepsilon(A_M) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_h^D} |S_h \lambda| \sum_{\substack{\boldsymbol{t} \geq \rho\lambda, \\ \|\boldsymbol{t}\| \leq \rho D}} \sum_{X \in [\mathbb{X}]_{\boldsymbol{t}}^\ell} \frac{1}{|A_M(X)|} [\delta(a_\lambda, X) \geq \varepsilon] .$$

Шаг 2. Пусть $X \in [\mathbb{X}]_{\boldsymbol{t}}^\ell$. Заметим, что алгоритм $a_{\boldsymbol{d}} \in A_M(X)$ тогда и только тогда, когда $\rho\boldsymbol{d} \leq \boldsymbol{t}$. Следовательно, $|A_M(X)| = (\lfloor t_1/\rho \rfloor + 1)(\lfloor t_2/\rho \rfloor + 1) \ldots (\lfloor t_h/\rho \rfloor + 1)$. Обозначим $T(\boldsymbol{v}) = \prod_j (v_j + 1)$. Тогда $|A(X)| = T\big(\lfloor \boldsymbol{t}/\rho \rfloor\big)$.

Шаг 3. Обозначим через $s = |U_1 \cap X|$ число объектов из U_1, лежащих в обучении. Тогда $\delta(a_\lambda, X) = \frac{m - s + \rho|\lambda|}{k} - \frac{s}{\ell}$, и условие $\delta(a_\lambda, X) \geq \varepsilon$ запишется в виде $s \leq \frac{\ell}{L}[m + \rho|\lambda| - \varepsilon k] \equiv s_0$. Множество

всех разбиений из $[\mathbb{X}]^\ell_{\boldsymbol{t}}$ с фиксированным параметром s обозначим через $[\mathbb{X}]^\ell_{\boldsymbol{t},s}$. Тогда

$$Q_\varepsilon(A_M) = \frac{1}{C^\ell_L} \sum_{\lambda\in Y_h^D} |S_h\lambda| \sum_{\substack{\boldsymbol{t}\geq\rho\lambda,\\ \|\boldsymbol{t}\|\leq\rho D}} \frac{1}{T\big(\lfloor \boldsymbol{t}/\rho\rfloor\big)} \sum_{s=0}^{s_0} \big|[\mathbb{X}]^\ell_{\boldsymbol{t},s}\big|.$$

Шаг 4. Вычислим мощность множества $[\mathbb{X}]^\ell_{\boldsymbol{t},s}$.

Введем обозначения $\ell' = \ell - \sum_{j=1}^h [t_j \neq \rho D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$. Тогда простое комбинаторное вычисление показывает, что $\big|[\mathbb{X}]^\ell_{\boldsymbol{t},s}\big| = C^s_m C^{k'-s}_{L'-m}$. Следовательно,

$$Q_\varepsilon(A_M) = \frac{1}{C^\ell_L} \sum_{\lambda\in Y_h^D} |S_h\lambda| \sum_{\substack{\boldsymbol{t}\geq\rho\lambda,\\ \|\boldsymbol{t}\|\leq\rho D}} \frac{1}{T\big(\lfloor \boldsymbol{t}/\rho\rfloor\big)} \sum_{s=0}^{s_0} C^s_m C^{k'-s}_{L'-m}.$$

Напомним, что $H^{\ell',m}_{L'}(z) = \sum\limits_{s=0}^{\lfloor z\rfloor} \frac{C^s_m C^{\ell'-s}_{L'-m}}{C^{\ell'}_{L'}}$ — функция гипергеометрического распределения [**?**]. Тогда

$$Q_\varepsilon(A_M) = \sum_{\lambda\in Y_h^D} \sum_{\substack{\boldsymbol{t}\geq\rho\lambda,\\ \|\boldsymbol{t}\|\leq\rho D}} \frac{|S_h\lambda|}{T\big(\lfloor \boldsymbol{t}/\rho\rfloor\big)} \frac{C^{\ell'}_{L'}}{C^\ell_L} H^{\ell',m}_{L'}(s_0).$$

■

Теорема 14. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к разреженной унимодальной сетке* $\ddot{A} = \{a_{\boldsymbol{d}}\}$ *размерности* h, $\|d\| \leq D$, *дается выражением:*

$$\begin{aligned} Q_\varepsilon(A) &= \sum_{\lambda\in Y_h^D} \sum_{\substack{\boldsymbol{t}\geq\rho\lambda,\\ \|\boldsymbol{t}\|\leq\rho D}} \sum_{\substack{\boldsymbol{t'}\geq 0,\\ \|\boldsymbol{t'}\|\leq\rho D}} \mathbb{S}(\lambda,\boldsymbol{t},\boldsymbol{t'}); \\ \mathbb{S}(\lambda,\boldsymbol{t},\boldsymbol{t'}) &= \frac{|S_h\lambda|\cdot 2^{|\lambda>0|}}{T\big(\lfloor \boldsymbol{t}/\rho\rfloor + \lfloor \boldsymbol{t'}/\rho\rfloor\big)} \frac{C^{\ell'}_{L'}}{C^\ell_L} H^{\ell',m}_{L'}(s_0), \end{aligned} \tag{21}$$

где $\ell' = \ell - \sum_{j=1}^{h}\big([t_j \neq \rho D] + [t'_j \neq \rho D]\big)$, $k' = k - |\boldsymbol{t}| - |\boldsymbol{t'}|$, *а остальные обозначения совпадают с обозначениями теоремы 13.*

Доказательство.

Шаг 1. Выберем в качестве представителя a_λ орбиты ω_λ алгоритм, не допускающий ошибок на множестве $Y = \bigcup_{j=1}^{h} Y_j$. Этого можно добиться, взяв произвольный $a_{\boldsymbol{d}} \in \omega_\lambda$ и поменяв знаки у всех $d_j < 0$ с помощью транспозиции z_j.

Введя обозначения $\boldsymbol{t}$ и $[\mathbb{X}]^\ell_{\boldsymbol{t}}$ так же, как и на первом шаге вывода формулы для монотонной сетки, получим

$$Q_\varepsilon(\ddot{A}) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_h^D} |S_h\lambda| \cdot 2^{|\lambda>0|} \sum_{\substack{\boldsymbol{t} \geq \rho\lambda, \\ \|\boldsymbol{t}\| \leq \rho D}} \sum_{X \in [\mathbb{X}]^\ell_{\boldsymbol{t}}} \frac{1}{|\ddot{A}(X)|} \left[\delta(a_\lambda, X) \geq \varepsilon\right].$$

Шаг 2. Обозначим с помощью t'_j максимальный индекс из $\{0, \ldots, \rho D\}$, при котором все объекты $\{y_j^1, \ldots, y_j^{t'_j}\}$ содержатся в $\bar{X}$, а $y_j^{t'_j+1}$, при его наличии, лежит в X. Положим $\boldsymbol{t'} = \{t'_j\}_{j=1}^h$. Заметим, что вектор $\boldsymbol{t'}$ играет для набора $\{Y_j\}$ ту же роль, что $\boldsymbol{t}$ для $\{X_j\}$. Обозначим через $[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'}}$ множество разбиений с фиксированными параметрами $\boldsymbol{t}$ и $\boldsymbol{t'}$.

Пусть $X \in [\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'}}$. Заметим, что $[a_{\boldsymbol{d}} \in \ddot{A}(X)] = [-\boldsymbol{t'} \leq \rho\boldsymbol{d} \leq \boldsymbol{t}]$. Следовательно, $|\ddot{A}(X)| = T\big(\lfloor \boldsymbol{t}/\rho \rfloor + \lfloor \boldsymbol{t'}/\rho \rfloor\big)$.

Шаг 3. Обозначим через $s = |U_1 \cap X|$ число объектов из U_1, лежащих в обучении. Пусть $s_0 \equiv \frac{\ell}{L}[m + \rho|\lambda| - \varepsilon k]$. Повторяя рассуждения аналогичного шага доказательства для разреженной монотонной сетки получим

$$Q_\varepsilon(\ddot{A}) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_h^D} |S_h\lambda| \cdot 2^{|\lambda>0|} \sum_{\substack{\boldsymbol{t} \geq \lambda, \\ \|\boldsymbol{t}\| \leq D}} \sum_{\substack{\boldsymbol{t'} \geq 0, \\ \|\boldsymbol{t'}\| \leq D}} \frac{1}{T(\boldsymbol{t}, \boldsymbol{t'})} \sum_{s=0}^{s_0} \big|[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'},s}\big|.$$

Шаг 4. Посчитаем мощность множества $[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'},s}$.

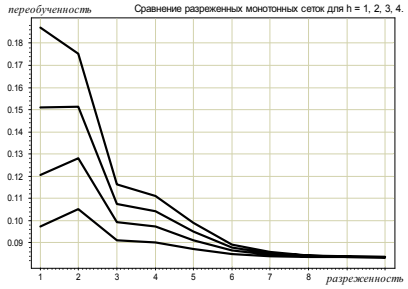


Рис. 10. Зависимость $Q_\varepsilon(\ddot{A})$ от разреженности ρ монотонной сетки при $L = 150$, $\ell = 90$, $\varepsilon = 0.05$, $D = 3$, $m = 5$, $h = 1, 2, 3, 4$.

Рис. 11. Зависимости вероятности переобучения $Q_\varepsilon(\ddot{A})$ для разреженной монотонной и унимодальной сеток от ρ при $L = 150$, $\ell = 90$, $\varepsilon = 0.05$, $D = 3$, $m = 5$, $h = 1(2), 2(4)$.

Обозначим $\ell' = \ell - \sum_{j=1}^{h}\big([t_j \neq \rho D] + [t'_j \neq \rho D]\big)$, $k' = k - |\boldsymbol{t}| - |\boldsymbol{t}'|$, $L' = \ell' + k'$. Тогда $\big|[\mathbb{X}]^{\ell}_{\boldsymbol{t},\boldsymbol{t}',s}\big| = C^s_m C^{k'-s}_{L'-m}$. Воспользовавшись определением функции гипергеометрического распределения получим:

$$Q_\varepsilon(A) = \sum_{\lambda \in Y_h^D} \sum_{\substack{\boldsymbol{t} \geq \rho\lambda, \\ \|\boldsymbol{t}\| \leq \rho D}} \sum_{\substack{\boldsymbol{t}' \geq 0, \\ \|\boldsymbol{t}'\| \leq \rho D}} \frac{|S_h\lambda| \cdot 2^{|\lambda>0|}}{T\big(\lfloor \boldsymbol{t}/\rho \rfloor + \lfloor \boldsymbol{t}'/\rho \rfloor\big)} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0).$$

■

Приведем результаты численных расчетов, иллюстрирующих поведение вероятности переобучения монотонной и унимодальной разреженных сеток. Расчеты выполнены с помощью доказанных выше формул (20), (21).

На рис. 10 изображена зависимость вероятности переобучения h-мерной монотонной сетки при $h = 1, 2, 3, 4$ от разреженности ρ. При увеличении размерности вероятности переобучения также

возрастает. При увеличении разреженности ρ вероятность переобучения падает, и вскоре выходит на константу, соответствующую вероятности переобучения лучшего алгоритма семейства a_0. Это связано с тем, что с уменьшением плотности семейства возрастает роль *явления расслоения* [**?**, **?**].

На рис. 11 приведены результаты сравнения разреженных h-мерных унимодальных сеток с разреженными $2h$-мерными монотонными сетками при $h = 1$ и $h = 2$. Тонкая серая кривая соответствует вероятности переобучения для унимодальной сетки. Полученные результаты подтверждают гипотезу [**?**] о связи вероятности переобучения для унимодальных сеток с вероятностью переобучения монотонных сеток удвоенной размерности.

2.7. Один слой хэммингова шара

Напомним, что расстояние между алгоритмами $\rho(a, a')$ определялось как расстояние Хэмминга между их векторами ошибок:

$$\rho(a, a') = \sum_{x \in \mathbb{X}} |I(a, x) - I(a', x)|.$$

Определение 7. *Шаром алгоритмов $B_r(a_0)$ радиуса r назовем множество, заданное условием $B_r(a_0) = \{a \in \mathbb{A} \colon \rho(a, a_0) \leq r\}$. Алгоритм a_0 будем называть* центром шара.

В отличие от рассмотренных в предыдущих разделах модельных семейств алгоритмов, хеммингов шар алгоритмов не имеет явного аналога среди реальных семейств алгоритмов. Однако, изучение этого модельного семейства, а также различных его подмножеств, представляет значительный теоретический интерес, поскольку хэммингов шар является максимально связным множеством булевых векторов. Одновременно с этим он состоит из огромного числа алгоритмов, значительно расслоеных по числу ошибок. Это делает хэммингов шар привлекательным примером для изучения влияния эффектов сходства и расслоения на вероятность переобучения.

Шары алгоритмов были подробно изучены в [**?**]. В частности, были получены точные формулы вероятности переобучения для рандомизированного метода минимизации эмпирического риска. При этом, поскольку эти модельные семейства обладают определенными симметриями, использовалась рассмотренная выше техника, основанная на действии группы симметрии $\mathrm{Sym}(A)$ на множестве алгоритмов и разбиений выборки.

Ниже мы рассматриваем следующее множество алгоритмов:

$$B_r^m(a_0) = \{a \in \mathbb{A} \colon n(a, \mathbb{X}) = m, \text{ и } \rho(a, a_0) \leq r\}.$$

Данное множество получается сечением шара $B_r(a_0)$ слоем алгоритмов с m ошибками. Следовательно, можно рассматривать его как пример наиболее «плотного» расположения алгоритмов внутри слоя с фиксированным числом ошибок.

Ограничимся изучением случая $n(a_0, \mathbb{X}) = m$. Тогда количество алгоритмов в $B_r^m(a_0)$ дается выражением $\sum_{i=0}^{\lfloor r/2 \rfloor} C_m^i C_{L-m}^i$. Данная величина чрезвычайно быстро растет с увеличением радиуса r. Мы покажем, что несмотря на стремительное увеличение числа алгоритмов рост вероятности переобучения оказывается весьма незначительным.

Положим без ограничения общности, что алгоритм a_0 ошибается на первых m объектах генеральной выборки $\mathbb{X}$. В дальнейшем множество, состоящее из первых m объектов $\mathbb{X}$, мы будем обозначать X^m. Множество, состоящее из последних $L - m$ объектов мы будем обозначать X^{L-m}. Для удобства m-й слой алгоритмов — множество, состоящее из всех алгоритмов, допускающих ровно m ошибок на генеральной выборке, — будем обозначать A_m.

Лемма 8. *Группа $S_m \times S_{L-m}$, где S_m и S_{L_m} — симметрические группы перестановок, действующие на множествах X^m и X^{L-m} соответственно, является подгруппой группы симметрий множества алгоритмов $B_r^m(a_0)$.*

Доказательство. Очевидно, что для $\pi \in S_m \times S_{L-m}$ справедливы следующие равенства: $n(a, X^m) = n\big(\pi(a), X^m\big)$ и $n(a, X^{L-m}) = n\big(\pi(a), X^{L-m}\big)$. Отсюда получаем:

$$\begin{aligned}\rho(a, a_0) &= m - n(a, X^m) + n(a, X^{L-m}) = \\ &= m - n\big(\pi(a), X^m\big) + n\big(\pi(a), X^{L-m}\big) = \rho\big(\pi(a), a_0\big).\end{aligned}$$

Таким образом, элементы группы $S_m \times S_{L-m}$ не меняют расстояния до центра шара $B_{r_0}(a_0)$.

Также заметим, что действие элементов симметрической группы S_L не меняет числа ошибок алгоритмов. Поскольку симметрическая группа действует на алгоритмы инъективно (у каждой перестановки π есть обратная ей), приходим к выводу, что $B_r^m(a_0) = \\ = \pi\big(B_r^m(a_0)\big)$. ■

Лемма 9. *Орбиты $\tau \in \Omega([\mathbb{X}]^\ell)$ действия группы $S_m \times S_{L-m}$ на множестве $[\mathbb{X}]^\ell$ индексированы параметром $i = |X \cap X^m| = \\ = n(a_0, X^\ell)$ — числом ошибок алгоритма a_0 на обучении. Мощность орбиты τ_i записывается в виде $|\tau_i| = C_L^\ell h_L^{\ell,m}(i)$.*

Доказательство. Первое утверждение леммы непосредственно следует из строения подгруппы симметрий, отмеченного в 8.

Мощность орбиты τ_i определяется числом способов независимо выбрать i объектов из X^m и $\ell - i$ объектов из X^{L-m}. Таким образом $|\tau_i| = C_L^\ell h_L^{\ell,m}(i)$. ■

Теорема 15. *Вероятность переобучения множества алгоритмов, получаемого сечением шара алгоритмов центральным m-слоем, дается в виде*

$$Q_\mu(\varepsilon, A) = H_L^{\ell,m}\big(s_d(\varepsilon) + \lfloor r/2 \rfloor\big),$$

где $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $H_L^{\ell,m}(s)$ — функция гипергеометрического распределения [?].

Доказательство. Заметим, что утверждение лемм 8 и 9 верно и для сечения шара центральной плоскостью. Поскольку все алгоритмы имеют равное число ошибок на полной выборке, применим следствие 1 из теоремы о разложении вероятности переобучения по орбитам разбиений выборки:

$$Q_\mu(\varepsilon, A) = \sum_{i=0}^{m} h_L^{\ell,m}(i) \left[\min_{a \in A} n(a, X_i) \leq \frac{\ell}{L}(m - \varepsilon k) \right].$$

Напомним, что по определению $i = |X \cap X^m|$. Пусть $r' = \left\lfloor \frac{r}{2} \right\rfloor$. Тогда выполнено следующее утверждение:

$$\min_{a \in A} n(a, X_i) = \begin{cases} 0, \text{ при } i \leq r', \\ i - r', \text{ при } i > r'. \end{cases}$$

Следовательно

$$Q_\mu(\varepsilon, A) = \sum_{i=0}^{\lfloor s_d(\varepsilon) \rfloor + r'} h_L^{\ell,m}(i) = H_L^{\ell,m}\big(s_d(\varepsilon) + \lfloor r/2 \rfloor\big),$$

где $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$. ■

На рис. 12 представлена зависимость точной оценки вероятности переобучения слоя шара, а также числа алгоритмов в семействе, от радиуса шара r. Видно, что за счёт значительной «плотности» данного семейства вероятность переобучения может оставаться на приемлемо низком уровне при мощности семейства порядка тысяч и относительно небольшой длине выборки $\ell = k = 100$. Заметим, что VC-оценки в этой ситуации вырождены и существенно превышают единицу.

2.8. Хэммингов шар

В прошлом разделе был рассмотрен пример наиболее «плотного» множества, наделенного свойством связности, но не обладаю-

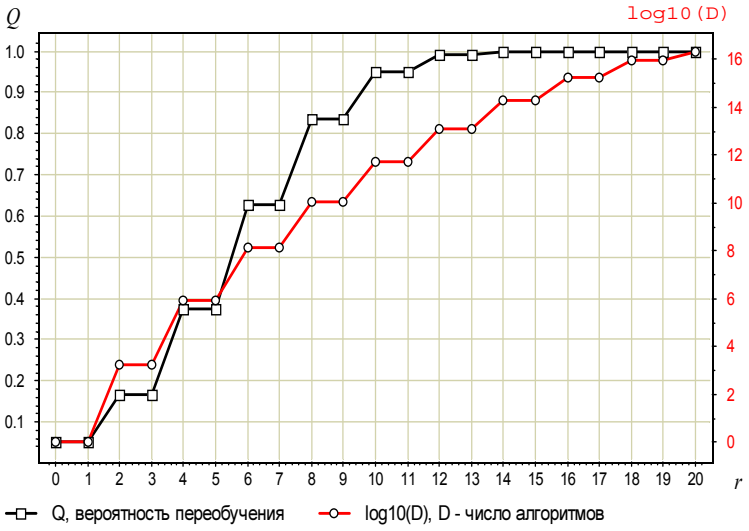


Рис. 12. Зависимость вероятности переобучения Q_ε и $\log_{10}|A|$ для слоя шара от радиуса шара r, при $\ell = k = 100$, $m = 10$, $\varepsilon = 0.05$.

щего свойством расслоения. Теперь мы изучим совместное влияние сходства и расслоения на вероятность переобучения. Хэммингов шар — наиболее подходящее для этого модельное семейство. Это пример наиболее «плотного» из одновременно связных и расслоенных семейств алгоритмов.

Пусть семейство алгоритмов A представляет собой хэммингов шар $B_{r_0}(a_0)$, где $n(a_0, \mathbb{X}) = m$. Пусть, без ограничения общности, алгоритм a_0 допускает ошибки на первых m объектах генеральной выборки $\mathbb{X}$.

Получим точную оценку вероятности переобучения для хэммингова шара алгоритмов.

Множество, состоящее из первых m объектов генеральной выборки $\mathbb{X}$, будем обозначать X^m, а множество, состоящее из последних $L - m$ объектов, будем обозначать X^{L-m}.

Лемма 10. *Группа $S_m \times S_{L-m}$, где S_m и S_{L-m} — симметрические группы перестановок, действующих на множествах X^m и X^{L-m} соответственно, является подгруппой группы симметрии семейства алгоритмов A.*

Доказательство. Очевидно, что для $\pi \in S_m \times S_{L-m}$ справедливы следующие равенства: $n(a, X^m) = n\big(\pi(a), X^m\big)$ и $n(a, X^{L-m}) = n\big(\pi(a), X^{L-m}\big)$. Отсюда получаем:

$$\begin{aligned}\rho(a, a_0) &= m - n(a, X^m) + n(a, X^{L-m}) = \\ &= m - n\big(\pi(a), X^m\big) + n\big(\pi(a), X^{L-m}\big) = \rho\big(\pi(a), a_0\big).\end{aligned}$$

Таким образом, элементы группы $S_m \times S_{L-m}$ не меняют расстояния до центра шара $B_{r_0}(a_0)$.

Тогда, если $a \in B_{r_0}(a_0)$, то и $\pi(a) \in B_{r_0}(a_0)$. Поскольку действие элементов симметрической группы на алгоритмы инъективно, приходим к выводу, что $B_{r_0}(a_0) = \pi\big(B_{r_0}(a_0)\big)$. ■

Лемма 11. *Орбитами действия группы $S_m \times S_{L-m}$ на множестве алгоритмов A являются пересечения слоев $m-r_0, \ldots, m+r_0$ со сферами радиусов $0, 1, \ldots, r_0$ и центрами в алгоритме a_0.*

Доказательство. Пусть алгоритм $a \in A_p$: $d(a, a_0) = r_1$. Мы установили, что в этом случае $d\big(\pi(a), a_0\big) = d(a, a_0) = r_1$. Также мы знаем, что действие перестановки на алгоритм не меняет числа его ошибок на $\mathbb{X}$. Таким образом, $\pi(a)$ также принадлежит пересечению p-го слоя со сферой радиуса r_1.

Осталось показать, что для любых $a_1, a_2 \in A_p$ таких, что $d(a_1, a_0) = d(a_2, a_0) = r_1$, найдётся перестановка $\pi \in S_m \times S_{L-m}$, при которой $\pi(a_1) = a_2$. Но это следует из того, что $n\left(a_1, X^m\right) = n\left(a_2, X^m\right)$ и $n\left(a_1, X^{L-m}\right) = n\left(a_2, X^{L-m}\right)$. Последний факт легко установить, выразив число ошибок, допускаемых алгоритмами a_1 и a_2 на множестве X^m, через m и r_1. ■

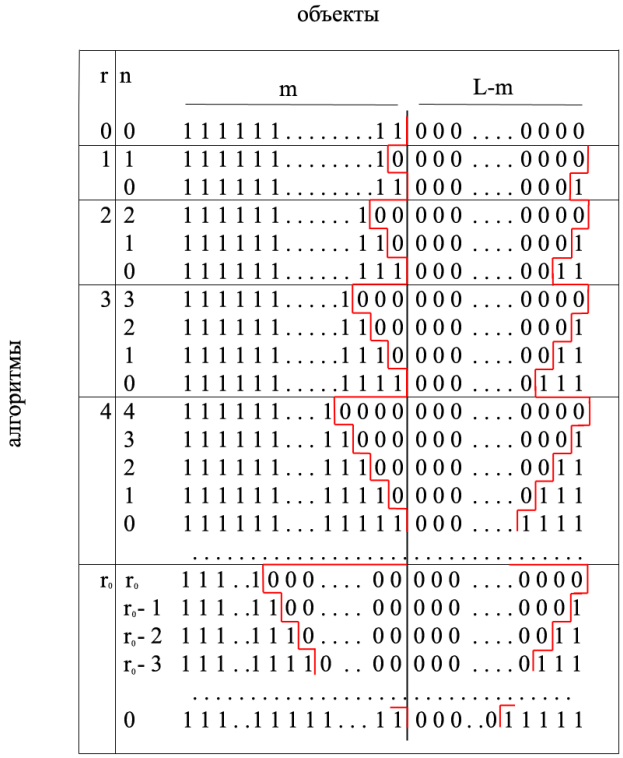


Рис. 13. Алгоритмы из орбит семейства A.

На рис. 13 для наглядности представлено по одному алгоритму из каждой орбиты семейства A. Пронумеруем орбиты двумя целочисленными индексами следующим образом: алгоритмы из орбиты $\mathrm{Orb}(r,n)$, $r=0,\ldots,r_0$, $n=0,\ldots,r$, принадлежат пересечению сферы радиуса r с центром в a_0 со слоем $r+m-2n$. Обратим внимание на то, что алгоритм $a_{(r,n)}$ из орбиты $\mathrm{Orb}(r,n)$ имеет $m-n$ единиц и n нулей на множестве X^m и $r-n$ единиц, $L-m-r+n$ нулей на множестве X^{L-m}. Всего в орбите $\mathrm{Orb}(r,n)$ содержится $C_m^n C_{L-m}^{r-n}$ алгоритмов.

В дальнейшем особую роль будет играть орбита $\mathrm{Orb}(r_0, r_0)$ — в неё входят алгоритмы шара A, допускающие наименьшее число ошибок на генеральной выборке $\mathbb{X}$. Вопреки закономерной аналогии с шаром в $\mathbb{R}^3$, для которого пересечение с касательной состоит из одной точки, нижний слой хэммингова шара состоит из $C_m^{r_0}$ алгоритмов.

Лемма 12. *Для любой обучающей выборки X и любого алгоритма $a \in A \setminus \mathrm{Orb}(r_0, r_0)$ выполнено:*

$$a \in A(X) \Leftrightarrow n(a, X) = 0.$$

Доказательство. Достаточность очевидна. Докажем необходимость.

Введём обозначения: $X^{\ell_1} = |X^m \cap X|$, $X^{\ell_2} = |X^{L-m} \cap X|$, $\ell_1 = |X^{\ell_1}|$, $\ell_2 = |X^{\ell_2}|$, $\ell = |X|$, $\ell = \ell_1 + \ell_2$, $X = X^{\ell_1} \cup X^{\ell_2}$.

Пусть $a \in A(X)$ и a принадлежит орбите $\mathrm{Orb}(r, n)$, отличной от $\mathrm{Orb}(r_0, r_0)$. Докажем, что алгоритм a не допускает ошибок на обучающей выборке X.

Начнем с рассмотрения случая $\ell_1 \leqslant r_0$. Алгоритмы орбиты $\mathrm{Orb}(\ell_1, \ell_1)$ имеют ровно ℓ_1 нулей на множестве X^m и не допускают ни одной ошибки на множестве X^{L-m}. Очевидно, существует алгоритм $a^{\ell_1} \in \mathrm{Orb}(\ell_1, \ell_1)$, такой что $n\left(a^{\ell_1}, X\right) = 0$.

В случае $\ell_1 > r_0$ в орбите $\mathrm{Orb}(r_0, r_0)$ существует алгоритм a^{r_0} для которого $n\left(a^{r_0}, X^{L-m}\right) = 0$ и $n\left(a^{r_0}, X^m\right) = \ell_1 - r_0$. Поскольку $n\left(a, X^m\right) \geqslant \ell_1 - n \geqslant \ell_1 - r_0 = n\left(a^{r_0}, X^m\right)$, то мы приходим к противоречию с тем, что $a \in A(X)$. ■

Таким образом, алгоритм не из $\mathrm{Orb}(r_0, r_0)$ может лежать в $A(X)$ только если он не допускает ошибок на обучающей выборке.

Следствие 5. *В ходе доказательства леммы 12 также установлено, что при $|X \cap X^m| > r_0$ алгоритмы из орбит, отличных от $\mathrm{Orb}(r_0, r_0)$, не могут попасть во множество $A(X)$.*

Лемма 13. *Для любого алгоритма $a \in \mathrm{Orb}(r_0, r_0)$ выполнено: если выборка $X \in [\mathbb{X}]^\ell$ такова, что $|X \cap X^m| \leqslant r_0$, то:*

$$a \in A(X) \Leftrightarrow n(a, X) = 0;$$

если X такова, что $|X \cap X^m| > r_0$, то:

$$a \in A(X) \Leftrightarrow n\left(a, X \cap X^m\right) = |X \cap X^m| - r_0.$$

Доказательство. Случай $|X \cap X^m| \leqslant r_0$ повторяет первую часть доказательства леммы 12. Рассмотрим случай $|X \cap X^m| > r_0$.

Поскольку в этом случае ни один алгоритм из множества A не допускает меньше $|X \cap X^m| - r_0$ ошибок на обучающей выборке X, а алгоритмы из $\mathrm{Orb}(r_0, r_0)$ не ошибаются на объектах множества X^{L-m}, то достаточность очевидна. Необходимость вытекает из того факта, что существует алгоритм $a \in \mathrm{Orb}(r_0, r_0)$, такой что $n(a, X) = n(a, X^m) = |X \cap X^m| - r_0$. ■

Лемма 14. *Орбитами действия группы $S_m \times S_{L-m}$ на множестве разбиений являются множества $\{X : |X \cap X^m| = i_0\}$, где $i_0 = \max(0, m-k), \ldots, \min(l, m)$.*

Доказательство. Утверждение леммы очевидным образом следует из определения действия группы $S_m \times S_{L-m}$ на множество разбиений. ■

Теорема 16 (Точная оценка для хэмингова шара)**.** *Пусть $n(a_0, \mathbb{X}) = m$. Рассмотрим шар алгоритмов $B_{r_0}(a_0)$. Тогда при обучении рандомизированным методом минимизации эмпирического риска и $r \leqslant \min(m, L-m)$ вероятность переобучения может быть записана в следующем виде:*

$$Q_\varepsilon(A) = \sum_{\substack{i=m-k; \\ i \geqslant 0}}^{r_0} h_L^{\ell,m}(i) \, \frac{\sum_{r=0}^{r_0} \sum_{n=0}^{r} S(n,r,i)}{\sum_{r=0}^{r_0} \sum_{n=0}^{r} S(n,r,i)} + \sum_{i=r_0+1}^{\lfloor s(\varepsilon) \rfloor} h_L^{\ell,m}(i),$$

где $S(n,r,i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n}$, $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k) + \frac{r_0 k}{L}$, $h_L^{\ell,m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}$.

Доказательство. Доказательство основано на применении теоремы 4, которая позволяет представить вероятность переобучения в виде:

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} \frac{|\omega|}{C_L^\ell} \sum_{\tau \in \Omega[\mathbb{X}]^\ell} |\{X \in \tau \colon a_\omega \in A(X)\}| \frac{\left[\delta(a_\omega, X_\tau) \geq \varepsilon\right]}{|A(X_\tau)|},$$

где $\Omega(A)$ — орбиты алгоритмов, а $\Omega[\mathbb{X}]^\ell$ — орбиты разбиений.

С учетом лемм 10, 11 и 14:

$$Q_\varepsilon(A) = \sum_{r=0}^{r_0} \sum_{n=0}^{r} \frac{C_m^n C_{L-m}^{r-n}}{C_L^l} \sum_{\substack{i=m-k; \\ i \geqslant 0}}^{\min(l,m)} \frac{S(i,r,n) \left[\delta\left(a_{(r,n)}, X_i\right) \geqslant \varepsilon\right]}{|A(X_i)|},$$

где $S(i,r,n) = \left|\{X \colon |X \cap X^m| = i,\ a_{(r,n)} \in A(X)\}\right|$, $a(r,n)$ — алгоритмы, представленные на рисунке 13, а X_i — произвольное разбиение, в обучающую выборку которого входят ровно i объектов из X^m.

Разобьем суммирование по орбитам разбиений (по i) на два слагаемых: $i \leq r_0$ и $i > r_0$.

$$\begin{aligned} Q_\varepsilon(A) = & \\ = & \sum_{r=0}^{r_0} \sum_{n=0}^{r} \frac{C_m^n C_{L-m}^{r-n}}{C_L^l} \sum_{\substack{i=m-k; \\ i \geqslant 0}}^{r_0} S(i,r,n) \frac{\left[\delta\left(a_{(r,n)}, X_i\right) \geqslant \varepsilon\right]}{|A(X_i)|} + \\ + & \sum_{r=0}^{r_0} \sum_{n=0}^{r} \frac{C_m^n C_{L-m}^{r-n}}{C_L^l} \sum_{i=r_0+1}^{\min(l,m)} S(i,r,n) \frac{\left[\delta\left(a_{(r,n)}, X_i\right) \geqslant \varepsilon\right]}{|A(X_i)|}. \end{aligned}$$

Из лемм 12 и 13, а также из следствия 5 следует, что первое слагаемое соответствует случаю выбора алгоритма, не допуска-

ющего ошибок на обучающей выборке X. Следствие 5 позволяет опустить суммирование по орбитам во втором слагаемом, поскольку при $i > r_0$ во множество $A(X_i)$ попадают только алгоритмы из $\mathrm{Orb}(r_0, r_0)$, допускающие в соответствии с леммой 13 ровно $i - r_0$ ошибок на X. С учетом этого:

$$Q_\varepsilon(A) = \\ = \sum_{r=0}^{r_0} \sum_{n=0}^{r} \frac{C_m^n C_{L-m}^{r-n}}{C_L^l} \sum_{\substack{i=m-k; \\ i \geqslant 0}}^{r_0} S\,(i,r,n)\, \frac{[r+m-2n \geqslant \varepsilon k]}{|A(X_i)|} + \\ + \frac{C_m^{r_0} C_{L-m}^0}{C_L^l} \sum_{i=r_0+1}^{\min(l,m)} S\,(i,r,n)\, \frac{[i \leqslant s(\varepsilon)]}{|A(X_i)|}. \quad (22)$$

Вычислим значение $S\,(i,r,n)$. В случае $i \leqslant r_0$ это число способов выбрать i из n объектов множества X^m и $l - i$ из $L - m - r + n$ объектов множества X^{L-m}, на которых не ошибается алгоритм $a_{(r,n)}$. В случае $i > r_0$ — число способов выбрать $i - r_0$ объектов из множества X^m, на которых алгоритм $a_{(r,n)}$ ошибается, и $l - i$ произвольных объектов из X^{L-m}. Итого:

$$S\,(i,r,n) = \begin{cases} C_n^i C_{L-m-r+n}^{l-i}, & i \leqslant r_0; \\ C_{m-r_0}^{i-r_0} C_{L-m}^{l-i}, & i > r_0. \end{cases}$$

Найдём значения $|A(X_i)|$. В случае $i > r_0$ в $A(X_i)$ попадают те алгоритмы $\mathrm{Orb}(r_0, r_0)$, которые ошибаются на множестве X^m $i - r_0$ раз. При $i \leqslant r_0$ из каждой орбиты во множество $A(X_i)$ попадают алгоритмы, не ошибающиеся ни на одном объекте множеств X^m и X^{L-m}. Получаем:

$$|A(X_i)| = \begin{cases} \sum\limits_{r=0}^{r_0} \sum\limits_{n=0}^{r} C_{m-i}^{n-i} C_{k-m+i}^{r-n}, & i \leqslant r_0; \\ C_i^{r_0}, & i > r_0. \end{cases}$$

Подстановка полученных результатов в (22) дает:

$$Q_\varepsilon(A) = \\ = \sum_{\substack{r=0 \\ r+m-2n\geqslant\varepsilon k}}^{r_0}\sum_{n=0}^{r} \frac{C_m^n C_{L-m}^{r-n}}{C_L^l} \sum_{\substack{i=m-k; \\ i\geqslant 0}}^{r_0} \frac{C_n^i C_{L-m-r+n}^{l-i}}{\sum\limits_{r_1=0}^{r_0}\sum\limits_{n_1=0}^{r_1} C_{m-i}^{n_1-i} C_{k-m+i}^{r_1-n_1}} + \\ + \frac{C_m^{r_0}}{C_L^l} \sum_{i=r_0+1}^{\min(l,m)} C_{m-r_0}^{i-r_0} C_{L-m}^{l-i} \frac{[i\leqslant s(\varepsilon)]}{C_i^{r_0}}. \quad (23)$$

Во втором слагаемом:

$$\frac{C_{m-r_0}^{i-r_0}}{C_i^{r_0}} = \frac{C_{m-r_0}^{m-i}}{C_i^{r_0}} = \frac{C_{i-(i-m+r_0)}^{r_0-(i-m+r_0)}}{C_i^{r_0}} = \frac{C_{r_0}^{i-m+r_0}}{C_i^{i-m+r_0}} = \frac{C_{r_0}^{m-i}}{C_i^{m-r_0}}.$$

Далее, $C_m^{r_0} C_{r_0}^{m-i} = C_m^i C_i^{r_0-m+i} = C_m^i C_i^{m-r_0}$. Подстановка полученных формул во второе слагаемое (23) дает:

$$\frac{C_m^{r_0}}{C_L^l} \sum_{i=r_0+1}^{\min(l,m)} C_{m-r_0}^{i-r_0} C_{L-m}^{l-i} \frac{[i\leqslant s(\varepsilon)]}{C_i^{r_0}} = \\ = \sum_{i=r_0+1}^{\min(l,m)} h_L^{l,m}(i)\,[i\leqslant s(\varepsilon)] = \sum_{i=r_0+1}^{\lfloor s(\varepsilon)\rfloor} h_L^{l,m}(i). \quad (24)$$

В первом слагаемом $C_m^n C_n^i = C_m^i C_{m-i}^{n-i}$, а

$$C_{L-m}^{r-n} C_{L-m-r+n}^{l-i} = C_{L-m}^{L-m-r+n} C_{L-m-r+n}^{l-i} = C_{L-m}^{l-i} C_{k-m+i}^{r-n}.$$

Заменив порядок суммирования, получим:

$$\sum_{\substack{r=0\\ r+m-2n\geqslant\varepsilon k}}^{r_0}\sum_{n=0}^{r}\frac{C_m^n C_{L-m}^{r-n}}{C_L^l}\sum_{\substack{i=m-k;\\ i\geqslant 0}}^{r_0}\frac{C_n^i C_{L-m-r+n}^{l-i}}{\sum\limits_{r_1=0}^{r_0}\sum\limits_{n_1=0}^{r_1}C_{m-i}^{n_1-i}C_{k-m+i}^{r_1-n_1}} =$$

$$= \sum_{\substack{i=m-k;\\ i\geqslant 0}}^{r_0} h_L^{l,m}(i)\frac{\sum_{r=0}^{r_0}\sum_{\substack{n=0\\ m+r-2n\geqslant\varepsilon k}}^{r} C_{m-i}^{n-i}C_{k-m+i}^{r-n}}{\sum_{r=0}^{r_0}\sum_{n=0}^{r}C_{m-i}^{n-i}C_{k-m+i}^{r-n}}. \quad (25)$$

Подстановка (24) и (25) в (23) завершает доказательство. ■

Замечание 1. *Впервые оценка, представленная в прошлом разделе (оценка для одного слоя хэммингова шара), была получена с помощью теоремы 4. Первый вариант ее доказательства занимал существенно больше места и был перегружен техническими деталями. В этом разделе приведено схожее доказательство оценки для шара, поскольку более изящного доказательства авторы не знают. Тем не менее интуиция подсказывает, что ключ к нему лежит в непосредственной работе с гипергеометрическим распределением в обход суммированию биномиальных коэффициентов.*

На рис. 14 представлены точные значения вкладов слоев шара в его вероятность переобучения. Видно, что несколько нижних слоев шара дают бóльшую часть вероятности переобучения. Возникает вопрос: нельзя ли приближать оценку шара оценкой для t его нижних слоев.

Замечание 2. *Поясним, почему график на рис. 14 имеет вид «гармошки». Понимание эффектов сходства и расслоения позволяют выдвинуть следующую гипотезу: вероятность переобучения расслоенного и связного множества алгоритмов может быть аппроксимирована вероятностью переобучения его подмножества, состоящего из существенно различных алгоритмов*

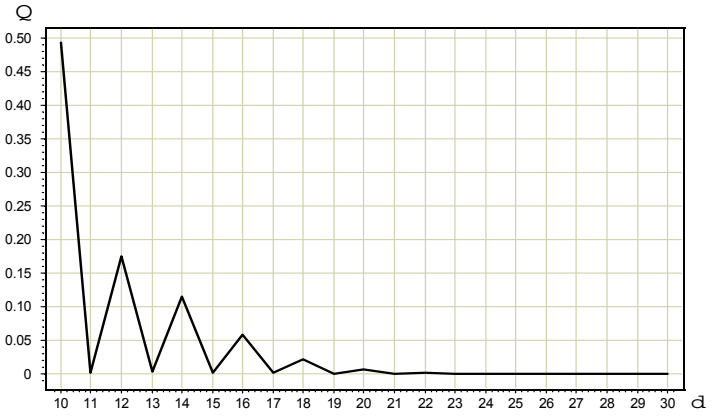


Рис. 14. Вклады d-слоёв шара в вероятность переобучения при $\ell = k = 100,\ m = 20,\ r_0 = 10,\ \varepsilon = 0.05$.

нижних слоев. Из рисунка 13 видно, что на внешней сфере шара представлены алгоритмы не из всех слоев: слои чередуются через один. Поскольку «существенно различные» алгоритмы лежат именно на внешней сфере, большие вклады в вероятность переобучения дают те слои шара, которые дают непустое пересечение с его внешней сферой, а именно — $m - r_0, m - r_0 + 2, \ldots$.

2.9. Нижние слои хэммингова шара

Теорема 17 (Точная оценка для нижних слоев хэммингова шара)**.** *Пусть $n(a_0, \mathbb{X}) = m$. Рассмотрим t нижних слоев шара алгоритмов $B_{r_0}(a_0)$. Тогда при обучении рандомизированным методом минимизации эмпирического риска и $r \leqslant \min(m, L - m)$ вероятность переобучения может быть записана в виде:*

$$Q_\varepsilon(A) = \sum_{\substack{i=m-k;\\ i\geqslant 0}}^{r_0} h_L^{\ell,m}(i)\, \frac{\sum_{r=0}^{r_0}\sum_{n=0}^{r} S'(n,r,i)}{\sum_{r=0}^{r_0}\sum_{n=0}^{r} S'(n,r,i)} + \sum_{i=r_0+1}^{\lfloor s(\varepsilon)\rfloor} h_L^{\ell,m}(i),$$

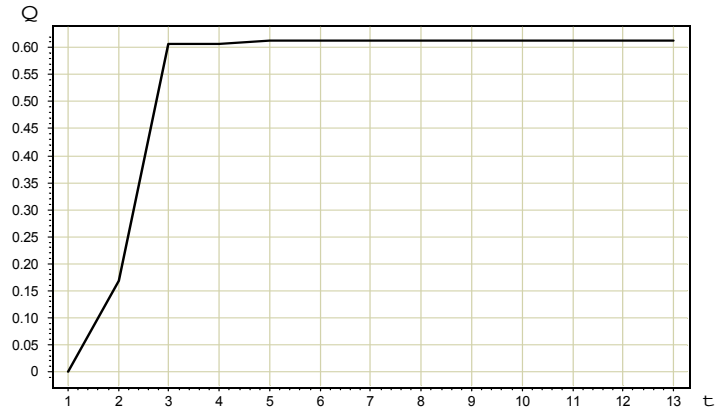


Рис. 15. Зависимость Q_ε от числа t нижних слоев шара, при $\ell = k = 100,\ m = 10,\ r_0 = 6,\ \varepsilon = 0.05$.

где $h_L^{\ell,m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}$, $S'(n,r,i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n}[r+r_0+1 \leqslant 2n+t]$, $s(\varepsilon) = \frac{\ell}{L}(m-\varepsilon k) + \frac{r_0 k}{L}$.

Доказательство повторяет доказательство теоремы 16. Леммы 10, 12 и 13 остаются справедливы для этого семейства алгоритмов. В лемме 11 множество слоев, в пересечении с которыми сферы дают орбиты алгоритмов, меняется на $m-r_0, \ldots, m-r_0+t-1$.

Основное отличие в ходе доказательства — при использовании теоремы 4 в начале доказательства множество суммируемых орбит алгоритмов сокращается добавлением проверочного множителя $[r + r_0 + 1 \leqslant 2n + t]$ после знаков суммирования по индексам r и n.

На рис. 15 представлена зависимость точной оценки вероятности переобучения для t нижних слоев шара от параметра t. Видно, что существенные скачки происходят лишь на первых нескольких слоях.

На рис. 16 представлены результаты приближения оценки шара t его нижними слоями. Черным цветом изображена оценка ша-

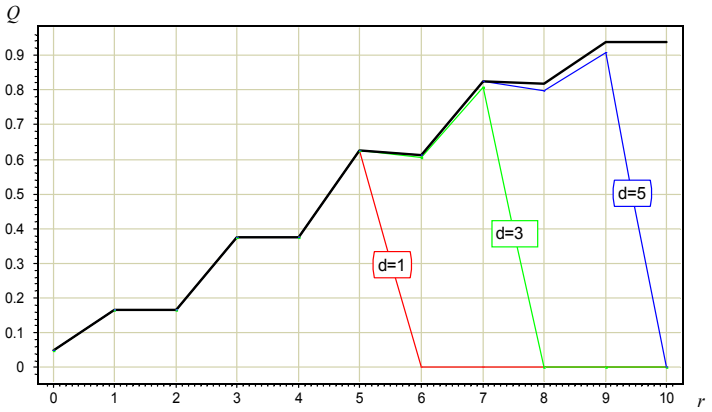


Рис. 16. Зависимость Q_ε от радиуса шара r для полного шара (верхняя кривая) и d его нижних слоёв, при $\ell = k = 100$, $m = 10,\ r_0 = 6,\ \varepsilon = 0.05$.

ра. Красным, зеленым и синим — оценки 1, 2 и 3 его нижних слоев соответственно. Падение к нулю оценок первых слоев шара объясняется уменьшением нижнего слоя шара с ростом его радиуса. В определенный момент количество ошибок m, допускаемое алгоритмами нижнего слоя шара, становится меньше εk. В этом случае переобучение невозможно.

Итак, в последних параграфах были исследованы хэммингов шар алгоритмов, его центральный слой и сечение несколькими его нижними слоями. На примере центрального слоя хэммингова шара мы еще раз продемонстрировали влияние эффекта связности на вероятность переобучения семейства. В то время, как вероятность переобучения множества, состоящего из случайных алгоритмов фиксированного слоя, в среднем экспоненциально растёт к единице с ростом его мощности (факт, который будет установлен в следующих разделах), вероятность переобучения центрального слоя хэммингова шара остается на достаточно низком уровне

даже при десятках тысяч алгоритмов в нем.

Также на примере всего шара и его нижних слоев удалось показать возможность приближения вероятности переобучения расслоенных семейств алгоритмов несколькими их нижними слоями.

3. Рандомизированные семейства алгоритмов

В предыдущем параграфе высказывалась гипотеза о приближении вероятности переобучения множества алгоритмов с помощью подмножества, состоящего из небольшого количества существенно различных алгоритмов, выбранных из нижних слоев. Ниже рассматриваются различные способы выбора таких подмножеств, и развивается теория позволяющая выводить соответствующие оценки вероятности переобучения.

3.1. Разреженные подмножества слоя

В одном из предыдущих параграфов была получена формула для вероятности переобучения сечения шара центральным слоем. Такие семейства являются наиболее «плотными» множествами алгоритмов, допускающими равное число ошибок. Ниже рассматривается другое подмножество слоя, в котором алгоритмы лишены сходства. Показывается, что вероятность переобучения в данном случае экспоненциально стремиться к единице с ростом числа алгоритмов.

Отметим, что задача построения множества алгоритмов фиксированной мощности с ограничением на минимальное попарное расстояние между алгоритмами хорошо изучена в теории кодов исправляющих ошибки. Однако методы построения кодов исправляющих ошибки накладывают существенные ограничения на параметры задачи: количество объектов в полной выборке и минимальные расстояния между алгоритмами. Предлагаемый ниже

подход позволит обойти задачу порождения несвязного множества алгоритмов заданной мощности.

Рассмотрим **упорядоченное** множество (т. е. вектор) алгоритмов $A = (a_1, \ldots, a_d) \subset \mathbb{A},\ d = |A|$, и группу $G = (S_L)^d$, состоящую из $(L!)^d$ элементов. Группа G действует на множество A, всевозможными способами переставляя ошибки разных алгоритмов независимо друг от друга.

Для формального определения действия G на A рассмотрим элемент группы $g = (\pi_1, \ldots, \pi_d) \in G$, где все $\pi_j \in S_L$. Он действует на множество A по правилу $gA = (\pi_1 a_1, \ldots, \pi_d a_d)$, где действие элементов группы S_L на $a \in \mathbb{A}$ определено перестановками объектов выборки так же, как и раньше.

Определим усредненный функционал вероятности переобучения:

$$\big\langle Q_\varepsilon(A)\big\rangle_G = \frac{1}{|G|}\sum_{g\in G} Q_\varepsilon(gA).$$

На самом деле, $\big\langle Q_\mu(\varepsilon, A)\big\rangle_G$ зависит уже не от самого множества алгоритмов A, а только от его профиля расслоения. Рассмотрим в качестве примера подмножество m-слоя, состоящее из d алгоритмов.

Теорема 18. *Пусть A — произвольное множество из D попарно-различных алгоритмов, каждый из которых допускает m ошибок на полной выборке. Тогда*

$$\big\langle Q_\varepsilon(A)\big\rangle_G = 1 - \Big(1 - H_L^{\ell,m}\left(s(\varepsilon)\right)\Big)^D, \tag{26}$$

где $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$.

Доказательство. Переставим знаки суммирования по $X \in [\mathbb{X}]^\ell$ и по $g \in G$ в функционале $\big\langle Q_\varepsilon(A)\big\rangle_G$:

$$\big\langle Q_\varepsilon(A)\big\rangle_G = \frac{1}{|G|}\frac{1}{C_L^\ell}\sum_{X\in[\mathbb{X}]^\ell}\sum_{g\in G}\sum_{a\in(gA)(X)}\frac{\big[\delta(a,X)\geq\varepsilon\big]}{|(gA)(X)|}.$$

Заметим, все слагаемые под знаком усреднения $\frac{1}{C_L^\ell}\sum_{X\in[\mathbb{X}]^\ell}$ равны друг другу. Поэтому, выбрав произвольного представителя $X \in [\mathbb{X}]^\ell$, запишем

$$\big\langle Q_\varepsilon(A)\big\rangle_G = \frac{1}{|G|}\sum_{g\in G}\sum_{a\in(gA)(X)}\frac{\big[\delta(a,X)\geq\varepsilon\big]}{|(gA)(X)|}. \tag{27}$$

Введем обозначение $s(\varepsilon) = \frac{\ell}{L}(m-\varepsilon k)$ и докажем, что для любого множества A, такого что все $a \in A$ имеют m ошибок на полной выборке, выполнено:

$$\sum_{a\in A(X)}\frac{\big[\delta(a,X)\geq\varepsilon\big]}{|A(X)|} = \Big[\min_{a\in A} n(a,X)\leq s(\varepsilon)\Big]. \tag{28}$$

Действительно, пусть $m_\ell = \min\limits_{a\in A} n(a,X)$. Тогда все алгоритмы из $A(X)$ имеют по m_ℓ ошибок на обучении и по $m_k = m - m_\ell$ ошибок на контроле. Следовательно, все $a \in A(X)$ имеют одинаковую переобученность $\delta(a,X) = \frac{m_k}{k} - \frac{m_\ell}{\ell}$. Отсюда немедленно получим $[\delta(a,X)\geq\varepsilon] = [m_\ell \leq s(\varepsilon)]$.

Подставляя (28) в (27), получим:

$$\begin{aligned}
\big\langle Q_\varepsilon(A)\big\rangle_G &= \frac{1}{|G|}\sum_{g\in G}\Big[\min_{a\in gA} n(a,X)\leq s(\varepsilon)\Big] = \\
&= \frac{1}{|G|}\sum_{g\in G}\Big[\exists a\in gA, n(a,X)\leq s(\varepsilon)\Big] = \\
&= 1 - \frac{1}{|G|}\sum_{g\in G}\Big[\forall a\in gA, n(a,X) > s(\varepsilon)\Big] = \\
&= 1 - \frac{1}{|G|}\sum_{g\in G}\prod_{i=1}^{D}\Big[n(\pi_i a_i, X) > s(\varepsilon)\Big].
\end{aligned}$$

Пользуясь тем, что группа G — декартово произведение групп, по-

лучим

$$\big\langle Q_\varepsilon(A)\big\rangle_G = 1 - \prod_{i=1}^{D} \frac{1}{|S_L|} \sum_{\pi_i \in S_L} \Big[n(\pi_i\, a_i, X) > s(\varepsilon) \Big].$$

Отметим, что все множители произведения $\prod_{i=1}^{D}$ равны друг другу. Зафиксируем произвольный $a \in A$, и пользуясь условием $n(\pi a, X) = n(a, \pi^{-1})X$, перепишем предыдущее выражение в виде

$$\big\langle Q_\varepsilon(A)\big\rangle_G = 1 - \Big(1 - \underbrace{\frac{1}{|S_L|} \sum_{\pi' \in S_L} \big[n(a, \pi' X) \leq s(\varepsilon) \big]}_{Q_\varepsilon(a)} \Big)^D,$$

где a — произвольный алгоритм с m ошибками на полной выборке.

Нетрудно заметить, что выделенное в предыдущей формуле выражение $Q_\varepsilon(a)$ равно $H_L^{\ell,m}\big(s(\varepsilon)\big)$ — вероятности переобучения вырожденного метода обучения, всегда возвращающего алгоритм a. ■

Замечание 3. *С увеличением числа алгоритмов в слое величина* $\Big(1 - H_L^{\ell,m}\big(s(\varepsilon)\big)\Big)^D$ *стремиться к нулю, а, следовательно, вероятность переобучения стремится к единице.*

На рисунке приведены результаты численного эксперимента, подтверждающего полученную выше формулу (26). По оси OX откладывается число алгоритмов, случайным образом выбранных из слоя с $m = 2$ ошибками. По оси OY откладывается разность между единицей и вероятностью переобучения для $\varepsilon = 0.04,\ L = 100,\ \ell = 60$. Гладкая кривая получена вычислением по формуле (26), тонкая кривая получена для одного случайно сгенерированного множества алгоритмов методом Монте-Карло по 50 тыс. разбиений.

Рис. 17. Рассеянный слой алгоритмов.

3.2. Разреженные подмножества слоя: случайный выбор без возвращения

В предыдущем параграфе было показано, что в отсутствии связности и расслоения вероятность переобучения экспоненциально быстро стремиться к единице. Отметим, что для доказательства этого факта нам пришлось отказаться от предположения попарной различности алгоритмов семейства. Этот шаг соответствует случаю выбора алгоритмов случайно *с возвращением*. Возможен и другой подход к задаче изучения вероятностей переобучения разреженных подмножеств слоя. Далее будет описана модель случайного выбора подмножества слоя *без возвращений*.

Мы будем рассматривать m-й слой алгоритмов A_m и случайно выбирать d различных алгоритмов из него. Требуется оценить

среднее значение вероятности переобучения таких случайных подмножеств m-го слоя:

$$\bar{Q}_\varepsilon(A_m, d) = \frac{1}{C_D^d} \sum_{\substack{A' \subset A_m: \\ |A'|=d}} Q_\varepsilon(A'),$$

где $D = |A_m| = C_L^m$ — число алгоритмов в m-м слое.

Теорема 19. *Среднее значение вероятности переобучения случайного подмножества A_m, состоящего из d различных алгоритмов, выражается следующим образом:*

$$\bar{Q}_\varepsilon(A_m, d) = 1 - \frac{C^d_{C_L^m \bar{H}_L^{\ell,m}(s(\varepsilon))}}{C^d_{C_L^m}},$$

где $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $\bar{H}_L^{\ell,m}(z) = \sum_{s=\lceil z \rceil}^{\ell} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — правый хвост гипергеометрического распределения.

Доказательство.

$$\bar{Q}_\varepsilon(A_m, d) = \frac{1}{C_D^d} \sum_{\substack{A' \subset A_m: \\ |A'|=d}} \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \frac{\sum_{a \in A'(X)} [\delta(a, X) \geqslant \varepsilon]}{|A'(X)|},$$

где $A'(X) = \arg\min\limits_{a \in A'} n(a, X)$. Несложно заметить, что в слое

$$\sum_{a \in A'(X)} [\delta(a, X) \geqslant \varepsilon] = \big|A'(X)\big| [\delta(a_0, X) \geqslant \varepsilon]$$

для произвольного $a_0 \in A'(X)$.

Таким образом,

$$\bar{Q}_\varepsilon(A_m, d) = \frac{1}{C_D^d} \sum_{\substack{A' \subset A_m: \\ |A'|=d}} \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \big[\exists a \in A' \colon n(a, X) \leqslant s(\varepsilon)\big].$$

Переставим местами знаки суммирования:

$$\bar{Q}_\varepsilon(A_m, d) = \\ = \frac{1}{C_D^d}\frac{1}{C_L^\ell}\sum_{X\in[\mathbb{X}]^\ell}\sum_{\substack{A'\subset A_m:\\ |A'|=d}}\bigl[\exists a\in A'\colon n(a,X)\leqslant s(\varepsilon)\bigr] = \\ = \frac{1}{C_D^d}\frac{1}{C_L^\ell}\sum_{X\in[\mathbb{X}]^\ell}F(X,A_m,d,\varepsilon).$$

Очевидно, что $F(X_1, A_m, d, \varepsilon) = F(X_2, A_m, d, \varepsilon)$ для любых разбиений X_1 и X_2. Зафиксируем произвольное разбиение X_0 и опустим первый аргумент у функции F:

$$\begin{aligned}\bar{Q}_\varepsilon(A_m, d) &= \\ &= \frac{1}{C_D^d}\sum_{\substack{A'\subset A_m:\\ |A'|=d}}\bigl[\exists a\in A'\colon n(a,X_0)\leqslant s(\varepsilon)\bigr] = \\ &= \frac{1}{C_D^d}\sum_{\substack{A'\subset A_m:\\ |A'|=d}}\bigl(1-\bigl[\forall a\in A'\colon n(a,X_0)> s(\varepsilon)\bigr]\bigr) = \\ &= 1-\frac{1}{C_D^d}\sum_{\substack{A'\subset A_m:\\ |A'|=d}}\prod_{a\in A'}\bigl[n(a,X_0)> s(\varepsilon)\bigr]. \qquad (29)\end{aligned}$$

Рассмотрим D различных бинарных переменных вида $z_i = [n(a_i, X_0) > s(\varepsilon)]$, $a_i \in A_m$, $i = 1, \ldots, D$. Нас интересует сумма всевозможных произведений d различных из них. Очевидно, что это число способов выбрать d ненулевых из них. С учетом этого равенство (29) примет следующий вид:

$$\bar{Q}_\varepsilon(A_m, d) = 1 - \sum_{1\leqslant i_1<\cdots<i_d\leqslant D}\frac{z_{i_1}\ldots z_{i_d}}{C_D^d} = 1-\frac{C^d_{|\{a\in A_m\colon n(a,X)>s(\varepsilon)\}|}}{C_D^d},$$

что после несложных вычислений завершает доказательство. ■

Из теоремы следует достаточно очевидный, но интересный факт:

Следствие 6. *Для любого семейства A, лежащего в m-м слое булева куба A_m, если $|A| \geqslant C_L^m \bar{H}_L^{\ell,m}\big(s(\varepsilon)\big)$, то $Q_\mu(\varepsilon, A) = 1$.*

Доказательство. При $d \geqslant C_L^m \bar{H}_L^{\ell,m}\big(s(\varepsilon)\big)$ сумма в последней строчке неравенства (29) равна 0, поскольку в каждое из слагаемых в этом случае будет входить нулевой множитель. ■

Замечание 4. *Мы снова получили экспоненциальный рост вероятности переобучения к единице с ростом числа алгоритмов d. Это легко увидеть, исследуя асимптотику логарифма дроби с использованием формулы Стирлинга.*

Помимо исследования разреженных семейств слоя и роста их переобученности с числом алгоритмов в них, описанный в этом разделе подход позволяет также изучать вопрос о приближении вероятности переобучения произвольного семейства алгоритмов с помощью его разреженных подмножеств. Это становится возможным, благодаря учету структуры самого семейства алгоритмов A, а не только его профиля расслоения, как в прошлом разделе.

В следующем разделе подобная возможность исследуется для семейств, лежащих в одном слое — в этом случае результат получается особо просто.

3.3. Разреженные подмножества семейств, лежащих в слое

Продолжая предложенный подход, сформулируем следующую теорему, которая, в некотором смысле, обобщает теорему 19:

Теорема 20. *Пусть $A \subseteq A_m$, $|A| = D$, где A_m — m-й слой алгоритмов. Среднее значение вероятности переобучения случайного подмножества A, состоящего из d различных алгоритмов, выражается следующим образом:*

$$\bar{Q}_\varepsilon(A, d) = 1 - \frac{\sum_{X \in [\mathbb{X}]^\ell} C^d_{N(X,\varepsilon)}}{C^d_D C^\ell_L},$$

где $N(X, \varepsilon) = \sum_{a \in A} [n(a, X) > s(\varepsilon)]$, $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$.

Доказательство теоремы полностью повторяет доказательство теоремы 19.

С помощью формулы Стирлинга снова легко показать, что каждое из C^ℓ_L слагаемых вида $C^d_{N(X,\varepsilon)}/C^d_D$ с ростом d экспоненциально быстро стремится к 0 либо 1 (для тех разбиений X, для которых $N(X, \varepsilon) = D$). При $d = |A|$ мы в точности получаем вероятность переобучения всего семейства A: $\bar{Q}_\varepsilon(A, |A|) = Q_\varepsilon(A)$.

Итак, мы получили важный результат — *в пределах одного слоя булева куба вероятность переобучения подмножеств произвольных семейств алгоритмов в среднем чрезвычайно быстро сходится к вероятности переобучения самого семейства с ростом размера подмножеств.*

Следующая лемма показывает, что вероятность переобучения множества алгоритмов, лежащего в одном слое, может только увеличиваться при добавлении к нему новых алгоритмов из того же слоя.

Лемма 15. *Пусть $A \subset A_m$ и $a \in A_m \setminus A$. Тогда $Q_\varepsilon(A) \leqslant Q_\varepsilon(A \cup a)$.*

Лемма 15 вместе с теоремой 20 дают нам простой способ оценки вероятности переобучения семейства, лежащего в слое. Сначала нужно определить размер d подмножества данного семейства, достаточно большой для требуемой точности оценки. Затем необходимо равномерно и случайно выбрать d алгоритмов из семейства и вычислить вероятность переобучения полученного подмножества. Тем не менее, на практике открытыми остается ряд

вопросов. Например, нужно иметь возможность равномерно «вытягивать» алгоритмы из семейства, полную матрицу ошибок которого мы, очевидно, не знаем.

4. Профили расслоения и связности множества алгоритмов

Поведение функционала $Q_\varepsilon(A)$ существенным образом зависит от структуры множества алгоритмов A. В то же время очевидно, что далеко не все возможные множества алгоритмов $A \subset \mathbb{A}$ возникают при решении практических задач обучения по прецедентам.

Во-первых, большинство реальных семейств обладают свойством связности: для каждого алгоритма $a \in A$ найдутся другие алгоритмы $a' \in A$ такие, что векторы ошибок $\boldsymbol{a}$ и $\boldsymbol{a}'$ отличаются только на одном объекте [**?**]. Связные семейства порождаются, в частности, методами классификации с непрерывной по параметрам разделяющей поверхностью. К ним относятся линейные классификаторы, машины опорных векторов с непрерывными ядрами, нейронные сети с непрерывными функциями активации, решающие деревья с пороговыми условиями ветвления, и многие другие. Эксперименты **??** показали, что с уменьшением связности семейства алгоритмов вероятность переобучения существенно возрастает.

Другим свойством, характерным для реальных семейств алгоритмов, является *расслоение*. Под m-слоем алгоритмов $A_m \subset A$ будем понимать подмножество алгоритмов, допускающих ровно m ошибок на полной выборке. Тогда профиль расслоения $\Delta(A, m)$ множества алгоритмов A определим как зависимость количества алгоритмов в m-слое $|A_m|$ от номера слоя m:

$$\Delta(A, m) = |A_m| = |\{a \in A \colon n(a, \mathbb{X}) = m\}|.$$

Эксперименты [**?**, **?**] показывают, что для большинства применя-

емых на практике семейств алгоритмов A профиль расслоения $\Delta(A, m)$ имеет форму узкого пика, сконцентрированного в средних слоях $m \approx L/2$. В то же время, вероятность переобучения в значительной мере определяется нижними слоями — алгоритмами с наименьшим числом ошибок на полной выборке. Алгоритмы высоких слоев имеют ничтожно малую вероятность реализоваться в результате обучения, и потому оказывают незначительный вклад в вероятность переобучения.

Изучение профиля расслоения и эффекта связности представляется важным шагом на пути к получению универсальных формул вероятности переобучения.

4.1. Профиль r-связности множества алгоритмов

В данном параграфе мы определим профиль r-связности множества алгоритмов и изучим его свойства. В частности, для задач классификации на два класса будет доказана инвариантность профиля r-связности по отношению к произвольной смене меток целевых классов объектов.

Напомним, что *расстояние между алгоритмами* $\rho(a, a')$ определялось как расстояние Хэмминга между их векторами ошибок на полной выборке:

$$\rho(a, a') = \sum_{x \in \mathbb{X}} |I(a, x) - I(a', x)|.$$

Шаром $B_r(a, A)$ радиуса r с центром в алгоритме a_0 называлось следующее множество множества алгоритмов A:

$$\{B_r(a, A) = a' \in A \colon \rho(a, a') \leq r\}.$$

Определение 8. *Профилем r-связности множества алгоритмов A назовем функцию от параметра q, заданную выражением:*

$$\Theta_r(q, A) = \sum_{a \in A} [|B_r(a, A)| = q].$$

Значение $\Theta_r(q, A)$ соответствует числу алгоритмов $a \in A$, имеющих ровно q соседей в шаре $B_r(a, A)$.

Рассмотрим задачу классификации выборки X_L на два класса $Y = \{+1, -1\}$. Обозначим целевой класс объекта $x_i \in X_L$ через $y_i \in Y$.

Рассмотрим действие группы $S_2 = \{e, h\}$ на множестве Y, при котором неединичный элемент $h \in S_2$ действует сменой метки целевого класса на противоположную: $+1 \leftrightarrow -1$.

Рассмотрим группу $G = (S_2)^L$ и ее элемент $g \in G = \{h_1, h_2, \ldots, h_L\}$. Элемент g действует на генеральную выборку $\mathbb{X}$ с помощью смены целевых классов у тех объектов x_i, для которых $h_i \neq e$. Действие группы G на $\mathbb{X}$ естественным образом продолжается до действия на множестве всех алгоритмов $\mathbb{A}$. Действительно, на каждый алгоритм $a \in \mathbb{A}$ элемент $g \in G$ действует по правилу

$$I(ga, x_i) = \begin{cases} +I(a, x_i), & \text{при } h_i = e \\ -I(a, x_i), & \text{при } h_i \neq e. \end{cases} \tag{30}$$

Это соответствует инверсии вектора ошибок на тех объектах, у которых g меняет целевой класс на противоположный.

Лемма 16. *Действие определенной выше группы G является изометрией относительно хэммингова расстояния ρ между векторами ошибок: для всех $a, a' \in \mathbb{A}$ и любого $g \in G$*

$$\rho(a, a') = \rho(ga, ga').$$

Доказательство.

Используя определение (30), убеждаемся, что

$$|I(ga, x_i) - I(ga', x_i)| = |I(a, x_i) - I(a', x_i)|$$

при обоих возможных значениях g_i. Тогда

$$\rho(ga, ga') = \sum_{x_i \in \mathbb{X}} |I(ga, x_i) - I(ga', x_i)| = \\ = \sum_{x_i \in \mathbb{X}} |I(a, x_i) - I(a', x_i)| = \rho(a, a').$$

■

Из доказанной выше леммы немедленно следует, что для задач бинарной классификации профиль r-связности инвариантен к смене меток целевых классов:

для любого $g \in (S_2)^L$ выполнено $\Theta_r(q, A) = \Theta_r(q, gA)$.

Таким образом, связность является топологическим свойством семейства алгоритмов, зависящим лишь от взаимного расположения классифицируемых объектов в пространстве признаков, но не зависящим от меток их классов. Проиллюстрируем предыдущее утверждение на примере семейства линейных классификаторов.

Рассмотрим задачу классификации точек трехмерного пространства разделяющими плоскостями, проходящими через центр координат: $y = \text{sign}(\langle w, x \rangle)$. Тогда, с топологической точки зрения, множеством алгоритмов будет сфера S^2. Точки сферы соответствуют единичному направляющему вектору разделяющей плоскости. Противоположные точки сферы соответствуют одновременной смене классификации всех объектов на противоположную.

Зафиксируем выборку объектов X_L, каждому из которых приписан целевой класс. Это порождает раскраску всей сферы на области, соответствующие алгоритмам с равными векторами ошибок. Заметим, что алгоритмы различаются на одном объекте тогда и только тогда, когда у них есть общая граница на полученной карте. Следовательно, профиль 1-связности полностью определяется отношением соседства алгоритмов. Границы карты, в свою

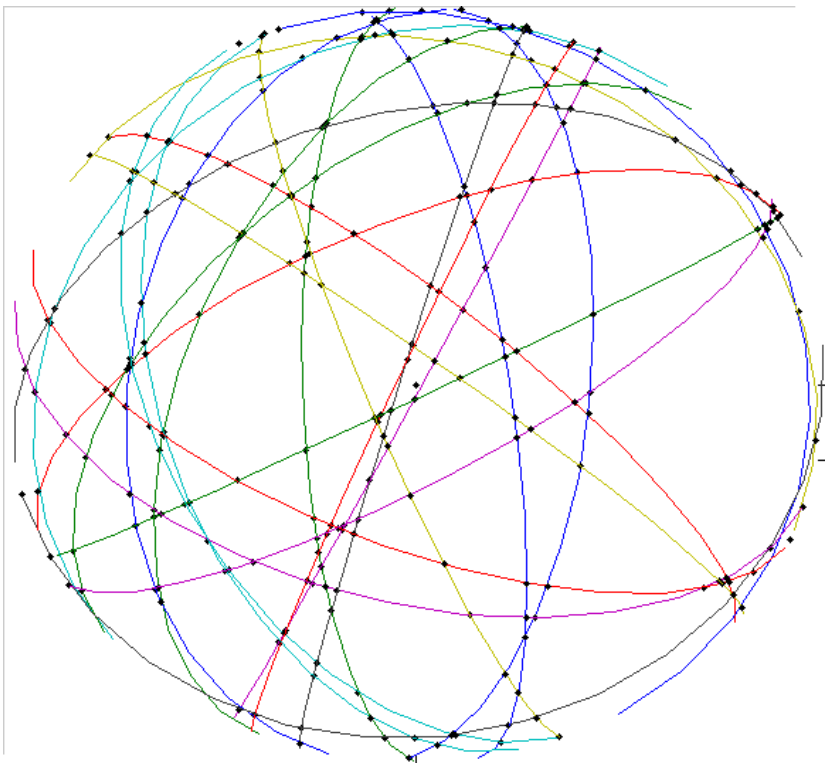


Рис. 18. Точки сферы — направляющие векторы линейной разделяющей гиперплоскости, грани графа на сфере — классы эквивалентных алгоритмов (с равными векторами ошибок); пары граней имеющих общую границу соответствуют алгоритмам, различающимся на одном объекте.

очередь, задаются условием $\langle w, x \rangle = 0$, инвариантным по отношению к меткам целевых классов. Следовательно, и профиль 1-связности не зависит от меток целевых классов.

Изучение топологической природы множества алгоритмов позволяет вывести ряд полезных свойств профиля r-связности, и, в частности, среднего числа связей алгоритмов. Для рассмотренного примера данную величину легко получить используя соотношение Эйлера между числом вершин, ребер и граней графа, реализованного на сфере.

Действительно, пусть точки выборки X_L находятся в общем положении. Тогда для каждой точки выборки $x \in X_L$ рассмотрим семейство разделяющих плоскостей $A_x \subset S^2$, проходящих через начало координат и заданную точку. На рассмотренной выше сфере алгоритмов данное множество является центральной окружностью, двойственной к рассматриваемой точке. Следовательно, граф границ алгоритмов получается с помощью сечения сферы центральными окружностями.

Каждая вершина графа получается как пересечение двух центральных окружностей. Следовательно, количество вершин $V = L(L-1)$. Из каждой вершины выходит ровно четыре ребра. Следовательно, количество ребер графа $E = 2L(L-1)$. Это позволяет определить количество граней графа:

$$V - E + S = \varphi(S^2),$$

где $\varphi(S^2) = 2$ — Эйлерова характеристика сферы. Следовательно, количество разделяющих гиперплоскостей в семействе равно $S = L(L-1) + 2$.

Среднее количество связей между $\frac{1}{|A|}\sum_{q=0}^{|A-1|} q \cdot \Theta_1(q, A)$ алгоритмами определяется отношением $2\,E/S$, т. е. асимптотически стремиться к 4 в рассматриваемом двумерном случае. В работе **??** доказывается более общее утверждение о том, что для семейства линейных классификаторов данная величина равна удвоенной размерности пространства признаков. Вопрос об изучении

профиля r-связности $\Theta_r(q, A)$ при $r \geq 2$, по всей видимости, еще не изучался.

4.2. Профиль расслоения-связности множества алгоритмов

Напомним, что профиль расслоения $\Delta(m, A)$ определялся как число алгоритмов $a \in A$, допускающих ровно m ошибок на объектах полной выборки:

$$\Delta(m, A) = \sum_{a \in A} [n(a, \mathbb{X}) = m].$$

Назовем профилем расслоения-связности $\Lambda_r(m, q, A)$ семейства A число алгоритмов $a \in A$ с m ошибками и q соседями в шаре радиуса r:

$$\Lambda_r(m, q, A) = \big|\{a \in A \colon |B_r(a, A)| = q \text{ и } n(a, \mathbb{X}) = m\}\big|.$$

Известны общие верхние оценки вероятности переобучения [**?**], использующие профиль расслоения-связности при $r = 1$. Вместе с тем было известно, что профиль расслоения-связности обладает рядом интересных свойств. В частности, для семейства линейных классификаторов было экспериментально показано, что профиль расслоениясвязности приближенно раскладывается в произведение двух функций, одна из которых зависит от числа ошибок алгоритма на полной выборке, а вторая — только от числа связей алгоритма.

Ниже мы приводим точную формулировку и доказательство данной гипотезы. Оказывается, точное равенство выполняется после усреднения профиля расслоениясвязности по действию группы $G = (S_2)^L$ всевозможных смен целевых классов объектов.

Лемма 17 (О наиболее вероятном профиле расслоения)**.**

$$\frac{1}{|G|} \sum_{g \in G} \Delta(m, gA) = \frac{C_L^m}{2^L} |A|.$$

Данное утверждение означает, что биномиальное распределение $P(m) = \frac{C_L^m}{2^L}$ задает статистически наиболее вероятное распределение числа алгоритмов по количеству ошибок. Необходимо учитывать, что для реальных задачах данное распределение будет иным. Смещение фактического распределения алгоритмов по числу ошибок относительно биномиального можно использовать в качестве меры информативности («удачности») использования данного семейства A для описания свойств конкретной выборки данных $\mathbb{X}$.

Доказательство.

Переставим знаки суммирования по $g \in G$ и по $a' \in A$:

$$\sum_{g \in G} \Delta(m, gA) = \sum_{g \in G} \sum_{a \in gA} [n(a, \mathbb{X}) = m] = \sum_{a' \in A} \sum_{g \in G} [n(ga', \mathbb{X}) = m].$$

Теперь осталось воспользоваться тем, что усредненный профиль расслоения, записанный для фиксированного алгоритма a', задается биномиальным распределением:

$$\text{для всех } a' \in A_m \text{ выполнено: } \frac{1}{|G|} \sum_{g \in G} [n(g\,a', \mathbb{X}) = m] = \frac{C_L^m}{2^L}.$$

■

Теорема 21 (О наиболее вероятном профиле расслоения-связности)**.**

$$\frac{1}{|G|} \sum_{g \in G} \Lambda_r(m, q, gA) = \frac{C_L^m}{2^L} \cdot \Theta_r(q, A).$$

Доказательство этой теоремы непосредственно следует из инвариантности профиля r-связности относительно действия группы G и леммы 17 о наиболее вероятном профиле расслоения.

Доказательство.

Обозначим оператор усреднения по действию группы G через $\mathbf{E}_G = \frac{1}{|G|}\sum_{g\in G}\cdot$

$$\begin{aligned}
&\mathbf{E}_G\Lambda_r(m,q,gA) = \\
&\quad = \mathbf{E}_G \sum_{a'\in A} [|B_r(ga',gA)| = q][n(ga',\mathbb{X}) = m] = \\
&\quad = \sum_{a'\in A} \mathbf{E}_G[|B_r(a',A)| = q][n(ga',\mathbb{X}) = m] = \\
&\quad = \sum_{a'\in A} \Big([|B_r(a',A)| = q]\cdot \mathbf{E}_G[n(ga',\mathbb{X}) = m]\Big) = \\
&\quad = \frac{C_L^m}{2^L}\cdot\Theta_r(q,A).
\end{aligned}$$

■

В работах [**?**] экспериментально установлено, что без усреднения по действию группы G аналогичное равенство выполнено приближенно:

$$\Lambda_r(m,q,A) \approx \frac{1}{|A|}\Delta(m,A)\cdot\Theta_r(q,A).$$

4.3. Экспериментальные результаты о профиле расслоения

Приведем экспериментальные результаты по сравнению наиболее вероятного профиля расслоения $\frac{C_L^m}{2^L}$ с реальными профилями, возникающими при классификации объектов плоскости линейными классификаторами.

На приведенных ниже рисунках рассматривалось четыре случая, отличающихся степенью линейной разделимости объектов выборки. Все сгенерированные выборки содержали 24 объекта (по 12 в каждом классе). Строились все возможные варианты линейной классификации моделями вида $y = \text{sign}(\langle w,x\rangle + w_0)$ и экспериментально вычислялось значения профиля расслоение

$\Delta(m, A)$. На графиках с профилями расслоения по оси X откладывается количество ошибок, по оси Y — доля алгоритмов с соответствующим уровнем ошибки. Более толстая кривая соответствует фактическому профилю расслоения, более тонкая — усредненному биномиальному.

Отметим, что смещение профиля расслоения относительно биномиального распределения можно использовать в качестве меры информативности множества алгоритмов. К сожалению, данный профиль является ненаблюдаемой величиной. Обычно поиск лучшего алгоритма обычно является итерационной процедурой, в ходе которой генерируется лишь некоторое подмножество алгоритмов из рассматриваемого семейства. Профиль расслоения наблюдаемого подмножества будет, очевидно, смещен будет смещен в сторону алгоритмов с малым количеством ошибок. Возможно, данное смещение удастся оценить, если провести калибровку: подать на вход алгоритму оптимизации шумовую выборку, полученную по исходном данным с помощью случайной смены меток целевых классов. Перспективным применением развитой выше теории представляется дальнейшее изучение критериев информативности основанных на наблюдаемом профиле распределения ошибок.

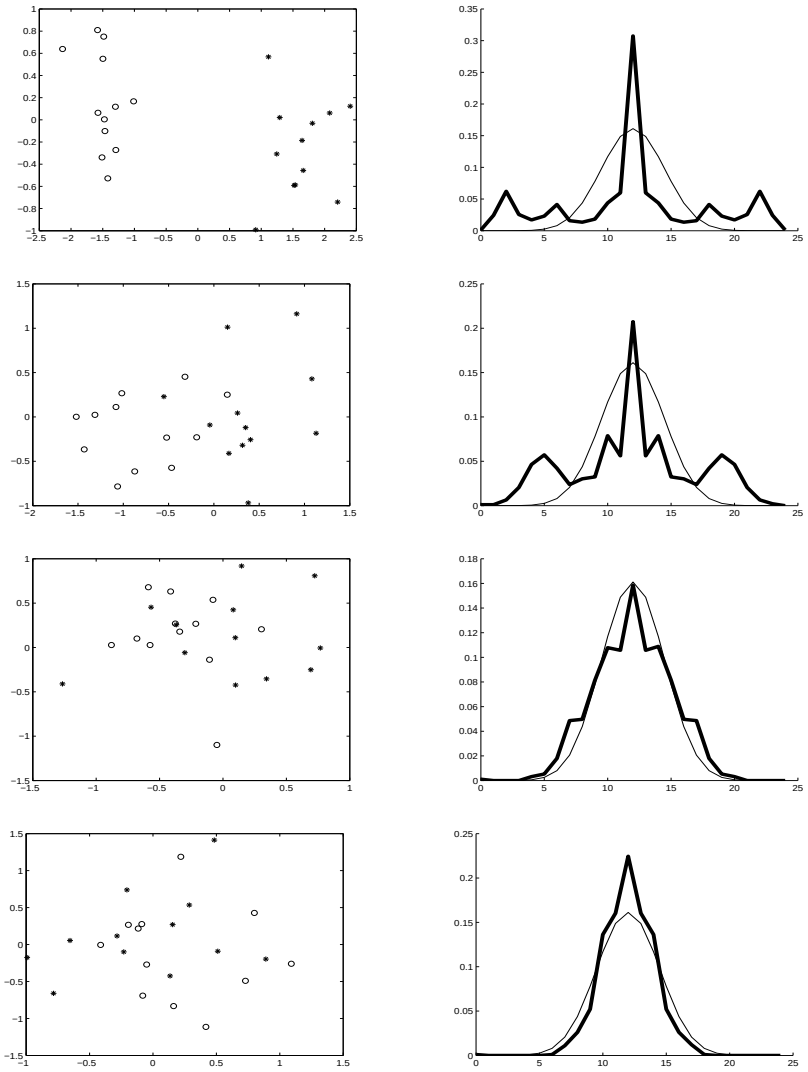


Рис. 19. Зависимость профиля расслоения от степени резделимости выборки.